eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Departamento de Lenguajes y Sistemas Informáticos

E-learning y la calibración de ítems de test: Teoría de Respuesta al Ítem versus calibración basada en juicios de expertos. Un estudio empírico

Rosa María Arruabarrena Santos

San Sebastián, julio de 2010

ISBN: 978-84-9860-513-6

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Departamento de Lenguajes y Sistemas Informáticos

E-learning y la calibración de ítems de test: Teoría de Respuesta al Ítem versus calibración basada en juicios de expertos. Un estudio empírico

Rosa María Arruabarrena Santos

San Sebastián, julio de 2010

Universidad del País Vasco Euskal Herriko Unibertsitatea

EXTRACTO/RESUMEN

Hezinet es un sistema hipermedia adaptativo (SHA) para el aprendizaje de euskera basado en la realización de actividades implementadas mediante ítems. Desde 2001 el Grupo de Hipermedia y Multimedia de la UPV/EHU viene realizando evaluaciones del sistema de cara a implementar mejoras para afinar la experiencia con el sistema de los diferentes usuarios. Uno de esos análisis llevó a observar que un problema a mejorar era la clasificación inicial de los nuevos alumnos añadiendo un módulo que administre pruebas de nivel mediante Tests Adaptativos Informatizados (TAI). Esta decisión provocará alguna modificación en los procesos de negocio asociados a la creación de los ítems que compondrán los tests, ya que éstos tienen que estar calibrados. Aunque en Hezinet los ítems se han calibrado siempre siguiendo la experiencia de los profesores, si se usa el método propuesto por la TRI, supone un concepto nuevo para los creadores de nuevo material que necesita de muestras muy numerosas y un procedimiento, aunque no complicado, si largo y tedioso, con muchas actividades alternativas, lo cual puede suponer un problema, por lo que se echa en falta un proceso definido a seguir. Es en este punto cuando surgen varias cuestiones: ¿Qué actividades tendría un proceso de calibración de un banco de ítems? ¿Se podrían emplear TAIs utilizando una calibración hecha por los profesores? ¿Qué recursos son necesarios para hacer dichas calibraciones? ¿Es equiparable una calibración hecha por profesores a la de la TRI? Si lo fuera, ¿cuándo es conveniente hacer cada una?

El trabajo de tesis se plantea con tres objetivos fundamentales: (1) formalizar una propuesta de proceso para la calibración de ítems utilizando la TRI, y establecer ciertas métricas asociadas para evaluar el consumo de recursos que conlleva; (2) crear una propuesta de proceso para la calibración de ítems utilizando juicios de expertos, así como aplicarles las mismas métricas que en el caso anterior; y (3) comparar las calibraciones de ítems obtenidas por el único rasgo que tienen en común ambas: la *dificultad*. Así mismo, contrastar el uso de recursos que tiene cada una de las opciones. Las dos propuestas enumeradas se pondrán en práctica para calibrar un conjunto de 252 ítems de partida que han sido entregados por *Zornotzako barnetegia* y que han servido de base para crear el sistema de clasificación de alumnos de *Boga,* la versión en Internet de Hezinet. Los ítems se calibrarán siguiendo cada uno de los procesos indicados utilizando métodos síncronos y asíncronos para la recogida de datos en cada uno de los casos. Como resultado de la experiencia se formalizará como propuesta uno de los procesos de calibración de ítems según la TRI y propondrá otro proceso para la calibración de ítems basada en los juicios de expertos. Además, se compararán los resultados obtenidos tanto en forma de calibración como de los recursos consumidos.

Índice general

Índice de ecuaciones

Índice de figuras

Índice de tablas

Acrónimos

A continuación, en orden alfabético, los acrónimos empleados a lo largo de esta memoria.

(1- β) – Potencia del contraste estadístico

1PL – One-parameter Logistic (Model)
Modelo logístico de un parámetro

2PL – Two-parameter Logistic (Model)
Modelo logístico de dos parámetros

3PL – Three-parameter Logistic (Model)
Modelo logístico de tres parámetros

3PL-TRI – Modelo logístico de tres parámetros de la Teoría de Respuesta al Ítem

a – Índice o parámetro de discriminación

α (nivel de significación) – Error de tipo I; es la probabilidad de obtener un valor de la estadística de prueba en la región crítica

ACER – *Australian Council for Educational Research*
Consejo Australiano de Investigación Educativa

AH – *Adaptive Hypermedia*

ANOVA – *ANalysis Of VAriance*

b – Índice o parámetro de dificultad del ítem

$\mathbf{b_i}$ –Parámetro de dificultad del ítem *i* en la calibración CT

β – Error de tipo II; es la probabilidad de aceptar la hipótesis nula cuando ésta es falsa

BOGA – Versión online de sistema Hezinet

BPMN – *Business Process Modeling Notation*
Notación de Modelado de Procesos de Negocios

BPMI – *Business Process Modeling Initiative*

c – Índice o parámetro de pseudoacierto

CCI – Curva Característica del Ítem

CE – Calibración de ítems empleando juicios otorgados por los expertos durante las pruebas de campo PE1 y PE2

CT – Calibración de ítems según el modelo 3PL de la TRI a partir valoraciones de los sujetos anónimos de las pruebas de campo PT1 y PT2

DELE – Diplomas de Español como Lengua Extranjera

D_i –Parámetro de dificultad del ítem *i* en la calibración CE

DILI – *Diploma Intermedio di Lingua Italiana*
Diploma de italiano como lengua extranjera, nivel intermedio

EAO – Enseñanza Asistida por Ordenador

EBAFLS – Banco Europeo de Ítems de Anclaje para la Evaluación de de Competencias en Lenguas Extranjeras

EGA – *Euskararen Gaitasun Agiria*,
Certificado de conocimientos básicos de euskara

ETS– *Educational Testing Service*
Servicio de Evaluación Educativa

FDI – Funcionamiento Diferencial de los Ítems

FII – Función de Información del Ítem

FIT – Función de Información del Test

GHyM – Grupo de investigación de Hipermedia y Multimedia

GQM – *Goal/Question/Metric*

H_0 – Hipótesis nula

H_1 – Hipótesis alternativa

HABE –*Helduen Alfabetatze eta Berreuskalduntzerako Erakundea*
Institución para la Alfabetización y Euskadunización de Adultos

Hezinet – Hezi + Net (Educar + Red)
Sistema hipermedia adaptativo multiusuario y multiplataforma para el aprendizaje del euskara

IC – Intervalo de Confianza

ICCS – *International Civic and Citizenship Study*
estudio internacional sobre educación cívica y ciudadanía

IEA – *Association for the Evaluation of Educational Achievement*

INECSE – Instituto Nacional de Evaluación y Calidad del Sistema Educativo

INES – *International iNdicators of Education Systems*

κ – índice Kappa-Fleiss de Cohen empleado para evaluar el nivel de concordancia entre varios observadores

MAP – Máxima A Posteriori (método de estimación bayesiana)

M.dif – Estadístico definido en CE para estimar el valor del parámetro Dificultad del ítem empleando juicios de expertos

M.est – Estadístico definido en CE para estimar el valor del parámetro Destreza lingüística trabajada por el ítem empleando juicios de expertos

MIR – Médico Interno Residente

NIER – Instituto de Investigación sobre Política Educativa de Japón

OCDE – *Organisation for Economic Co-operation and Development*
Organización para la Cooperación y el Desarrollo Económico

p – Nivel de significación estadístico

PIR – Psicólogo Interno Residente

PE1 (PE2) –Prueba de Campo 1 (2) con expertos

PDCA – *Plan/Do/Check/Act*
Planificar/Ejecutar/Analiza/Mejorar

PET – *Preliminary English Test*
Diploma de inglés como lengua extranjera, nivel intermedio

PIAAC – Programa para la evaluación internacional de las competencias de adultos

PIRLS – *Progress in Internacional Reading Literacy Study*
Estudio Internacional de Progreso en Comprensión Lectora

PISA – *Programme for International Student Assessment*
Programa para la Evaluación Internacional de Alumnos

PT1 (PT2) – Prueba de Campo 1 (2) con participantes anónimos de la calibración 3PL-TRI

RAND – *Research ANd Development corporation*

SHA – Sistema Hipermedia Adaptativo

SQuaRE – *Software product Quality Requirements and Evaluation*

STI – Sistema Tutor Inteligente

SweSAT – *Swedish Scholastic Aptitude Test*
pruebas de admisión a las universidades de suecia

TAI – Test Adaptativo Informatizado

TALIS – Teaching and Learning International Survey

TCT – Teoría Clásica de los Tests

TEDS-M – *Teacher EDucation Study in Mathematics*
estudio internacional que evalúa la formación inicial del profesorado de Matemáticas en educación primaria y secundaria obligatoria

TIMSS – *Trends in International Mathematics and Science Study*
Estudio Internacional de tendencias en Matemáticas y Ciencias

TRI – Teoría de Respuesta al Ítem

UPV/EHU – Universidad del País Vasco

θ – Nivel de habilidad o rasgo del alumno

Dedicatoria

A Tomás A. Pérez, Javier López-Cuadrado, Anaje Armendáriz,
José Ángel Vadillo, Silvia Sanz, Julián Gutiérrez y
al resto de personas de mi grupo de investigación
que han hecho posible que hoy estemos aquí.

A Itziar Irigoyen, Basilio Sierra, José Yurramendi y Endika Bengoetxea
por sus conocimientos estadísticos.

A Philippe Lopistéguy, Juan Manuel Pikatza e Imanol Usandizaga
por sus consejos en el campo de la Ingeniería del Software.

A mis compañeros y amigos, Arantxa, Carmen, Joserra, José, Juan y Mar,
por brindarme vuestro apoyo.

A Bingen, Libe, Iruri, Coro y Carmen, mi familia, por la paciencia
que habéis depositado en mí a lo largo de esta trayectoria.

A todas aquellas personas que han contribuido de alguna manera en
que este proyecto se haya finalmente hecho realidad.

PARTE 1
Introducción

*En la **Parte 1** se introduce el trabajo de tesis incluido en esta memoria. Se presenta una propuesta de procesos de negocio de calibración de ítems para realizar la ubicación en su correspondiente nivel a alumnos nuevos del sistema de e-learning de la lengua vasca Hezinet. Para cumplir los objetivos de esta parte, se han incluido dos capítulos.*

*El **capítulo 1** describe brevemente los objetivos de la tesis, el trabajo realizado y se ubica la investigación desarrollada en el contexto del grupo de investigación del que forma parte la autora. Ahí mismo se explica la estructura de la tesis con más detalle y se da una pequeña guía de lectura para el lector.*

*El **capítulo 2** se dedica a detallar Hezinet, el sistema hipermedia adaptativo para el aprendizaje del euskera que ha servido de punto de partida para el desarrollo de la investigación. Se comentan también las teorías en las que se basa el sistema y se describe el funcionamiento que tiene el mismo.*

Capítulo 1
Objetivos y contexto

El Grupo de Hipermedia y Multimedia (GHyM) de la Facultad de Informática de San Sebastián y ligado al Departamento de Lenguajes y Sistemas Informáticos de la Universidad del País Vasco (UPV/EHU) trabaja en el área del *e-learning* desde 1994, cuando empezó a diseñar el sistema hipermedia adaptativo Hipertutor. Después de realizar sobre él ciertas modificaciones, se obtuvo **Hezinet**, un sistema hipermedia adaptativo (SHA) para el aprendizaje de *euskera* basado en la realización de actividades implementadas mediante ítems (Pérez, 2000).

La *Institución para la Alfabetización y Euskadunización de Adultos* (HABE) del Gobierno Vasco lo adquirió y se ha instalado en más de 150 centros del País Vasco, incluidos locales municipales y *euskaltegis*, las academias homologadas por el gobierno autónomo vasco para la enseñanza a adultos de euskera. Fuera de la Comunidad Autónoma Vasca, se encuentra disponible también en algunos centros educativos y culturales ubicados en Cataluña, Madrid, Argentina, Chile, Estados Unidos, Francia, Inglaterra, México, Puerto Rico, Uruguay y Venezuela.

Desde 2001, el grupo de investigación se ha centrado en la evaluación del sistema de aprendizaje (Arruabarrena, Pérez et al., 2001) con el objetivo de realizar mejoras continuas sobre la aplicación y, en algunos casos, su evolución a una nueva versión, y generalmente con un nuevo nombre (como *Zibergela* o *Boga*). Para simplificar, y puesto que no es relevante distinguir cada uno de ellos, en esta memoria se utilizará para todas las versiones el nombre original de *Hezinet*.

Las mejoras han estado enfocadas a afinar la experiencia con el sistema de los diferentes usuarios que tiene: los *alumnos,* los *profesores-tutores* (o *instructores*), o los *creadores de material nuevo para el sistema.*

El usuario principal es el **alumno**, que interactúa con el sistema para obtener como resultado una adquisición de conocimiento, aprendizaje de nuevos conceptos. El **instructor** se encarga de asegurarse de que la experiencia del aprendizaje sea lo más fructífera posible, comprobando si existen dificultades que ralentizan el avance o revisando las decisiones del sistema por si fuera necesario reubicar o revisar ciertos conceptos y el sistema no lo hubiese detectado. Además, el **creador** de nuevo material puede generar material adicional para ser incluido dentro del sistema como métodos alternativos para mostrar un concepto (por ejemplo, usando un video o mediante un texto, una historia interactiva, un audio, etc.); o versiones distintas para explicarlo (como maneras distintas de explicar una misma idea usando un mismo medio); o nuevos ítems para evaluar un mismo conocimiento.

Tras los pertinentes estudios se observó que uno de los mayores cuellos de botella del sistema era la clasificación inicial de los nuevos alumnos, ya que la afluencia de alumnos se concentraba en ciertos momentos haciendo que ésta fuera lenta y gran consumidora de recursos. Hasta el momento el sistema administraba tests secuencialmente con ítems de dificultad creciente correspondientes a cursos consecutivos. Este método suponía que todos los alumnos tenían que ir aprobando todos los test de niveles inferiores hasta llegar al suyo, que sería aquél que no se superase.

Esta tarea debería ser innecesaria para los alumnos realmente noveles (su nivel es claramente el primero). Es más, tener que responder un test que seguro se va a suspender puede ser desmotivador. Los alumnos de niveles bajos, se pueden ver abrumados por la cantidad de ítems que no saben responder, algo directamente relacionado del tamaño de cada uno de los tests de nivel. Por el contrario, los alumos que se valoren de niveles altos pueden confiarse por la cantidad de ítems que saben responder correctamente frente a los que no, ya que, a medida que su nivel crece, la proporción de ítems que se responderá incorrectamente

decrece, siendo, además, las respuestas de los primeros tests irrelevantes para determinar su nivel. Otra posible consecuencia es la desmotivación que surge de la administración de ítems fáciles a este tipo de alumnos. Además, entre test y test, deben esperar por la valoración del instructor sobre cómo ha realizado el test anterior.

Para el instructor, ubicar a gran cantidad de alumnos le requiere valorar cada test realizado para determinar si se supera y decidir, en consecuencia, si es conveniente realizar el siguiente test en dificultad. Y todo ello en momentos en los que hay múltiples alumnos realizando tests, que esperan su resolución cada vez que acaban cada uno de ellos. Si, además, se pretende evitar que se realicen tantos tests, es necesaria una interacción personal con cada alumno (como una entrevista) que ayude al instructor a ubicarlo y reducir la carga reduciendo la cantidad de ítems a completar, lo cual supone aún mayor esfuerzo. Siempre en un periodo de tiempo muy breve.

El creador de material tiene que preocuparse por temas de sobreexposición. Esta es una preocupación fundamental de los responsables de la evaluación en contextos aplicados ya que, incluso cuando se decide aplicar tests convencionales, uno de los mayores obstáculos a la validez de los tests es que los evaluandos puedan conocer de antemano los ítems que se le van a administrar. Puesto que los test eran fijos, los alumnos podían conocer de antemano algunos de los ítems que forman parte del test, e incluso conocer las respuestas sin tener los conocimientos pertinentes. Es por ello que se deben renovar con cierta frecuencia los ítems incluidos en cada uno de los tests, de manera que el problema mencionado se reduzca.

La propuesta de mejora consiste en reemplazar el método de ingreso por otro de evaluación que genere *pruebas de nivel adaptativas* que se implementarían mediante Tests Adaptativos Informatizados (TAIs) de ingreso al sistema (López-Cuadrado, 2003; López-Cuadrado, Pérez et al., 2002).

Esta decisión supone que los ítems que se utilizarán deben estar calibrados, una nueva tarea sobre cada nuevo material que se reflejará en un alivio para los instructores durante la época de valoración de nuevos alumnos.

Aunque en Hezinet los ítems se han calibrado siempre siguiendo la experiencia de los profesores, la necesidad de implantar los TAIs, exige que a cada ítem se le estime, al menos, un nivel de dificultad. Si se usa el método propuesto por la TRI, supone un concepto nuevo para los creadores de nuevo material que necesita de muestras muy numerosas y un procedimiento, aunque no complicado, si largo y tedioso, con muchas actividades alternativas (López-Cuadrado, 2008), lo cual puede suponer un problema, por lo que se echa en falta un proceso definido a seguir.

Es en este punto cuando surgen varias cuestiones: ¿Qué actividades tendría un proceso de calibración de un banco de ítems? ¿Se podrían emplear TAIs utilizando una calibración hecha por los profesores? ¿Qué recursos son necesarios para hacer dichas calibraciones? ¿Es equiparable una calibración hecha por profesores a la de la TRI? Si lo fuera, ¿cuándo es conveniente hacer cada una?

La presente tesis resuelve las dudas planteadas y alivia alguno de los problemas comentados. A continuación se presentan los objetivos de la misma. Seguidamente se ofrece una guía de lectura de la misma según el perfil del lector.

1.1. Objetivos

El trabajo de tesis se plantea con tres objetivos fundamentales:

(1) formalizar una propuesta de proceso para la calibración de ítems utilizando la TRI, y establecer ciertas métricas asociadas para evaluar el consumo de recursos que conlleva; No se contempla un estudio exhaustivo de todas las alternativas existentes para hacer la calibración, sino que se busca un procedimiento probado que nos lleve a la obtención de una estimación aceptable de los parámetros de un banco de ítems.

(2) crear una propuesta de proceso para la calibración de ítems utilizando juicios de expertos, así como aplicarles las mismas métricas que en el caso anterior; y

(3) comparar las calibraciones de ítems obtenidas por el rasgo que tienen en común: la *dificultad*. Así mismo, contrastar el uso de recursos que tiene cada una de las opciones.

Las dos propuestas enumeradas se pondrán en práctica para calibrar un conjunto de 252 ítems de partida que han sido desarrollados y entregados por *Zornotzako barnetegia* y que han servido de base para crear el sistema de clasificación de alumnos de *Boga,* la versión en Internet de Hezinet. Los ítems se calibrarán siguiendo cada uno de los procesos indicados utilizando métodos síncronos y asíncronos para la recogida de datos en cada uno de los casos. Como resultado de la experiencia se formalizará como propuesta uno de los procesos de calibración de ítems según la TRI y propondrá otro proceso para la calibración de ítems basada en los juicios de expertos. Además, se compararán los resultados obtenidos tanto en forma de calibración como de los recursos consumidos.

1.2. Evolución y ámbitos de trabajo

Hezinet toma la *mejora continua* como paradigma para su evolución. Así y desde su concepción la aplicación Hezinet ha soportado varias evaluaciones. Concretamente, como consecuencia de una primera evaluación formativa la interfaz de Hezinet se modificó substancialmente y se le añadió alguna funcionalidad más al sistema antes de su lanzamiento al mercado (Pérez, López et al., 2000). Tanto su rápida expansión a nivel mundial tras su comercialización como su integración con otro tipo de aplicaciones relacionadas con la cultura vasca (Armendáriz, López-Cuadrado et al., 2004) avalan de manera informal la validez del sistema.

En esta situación, se realizó un estudio con vistas a identificar posibles áreas de mejora, y se elaboró un plan de evaluación sumativa que se recogió en (Arruabarrena, Pérez et al., 2001; Arruabarrena, Pérez et al., 2002). Específicamente, las evaluaciones sumativas planificadas fueron: (1) medir el nivel de amigabilidad que presentaba la interfaz, (2) concretar la incorporación o mejora de nuevas prestaciones del sistema, (3) evaluar el impacto afectivo de Hezinet sobre los distintos tipos de usuarios del sistema, (4) medir la efectividad del sistema en el progreso del aprendizaje de los distintos tipos de alumnos que trabajaron con y sin el sistema en los euskaltegis y (5) valorar la integración de la herramienta dentro del proceso de aprendizaje del euskera desarrollado en los euskaltegis.

Relacionado con el plan de evaluación de Hezinet propuesto, en (Villamañe, Gutiérrez et al., 2001) se presentó una comparación entre los resultados obtenidos en exámenes oficiales por alumnos de euskaltegis que habían empleado la herramienta como apoyo al aprendizaje del euskera y los que no. Desde el punto de vista de efectividad del sistema, indicar que los usuarios de Hezinet obtuvieron mejores resultados; y, desde el punto de vista afectivo, consideraron muy positiva la experiencia con el sistema.

Así mismo, en este estudio se observó que el primer curso de Hezinet, el de nivel más básico, era demasiado elevado para los alumnos que desconocían totalmente el idioma. En una versión posterior, se incorporó un nuevo nivel, el cero, especialmente adaptado a aquellos que tienen un conocimiento nulo del idioma.

En la actualidad este sistema se usa en tres ámbitos diferenciados cada una de ellos con una denominación distinta (Hezinet, Zibergela y Boga) aunque con el mismo núcleo operativo.

La versión pionera del sistema se denominó **Hezinet** (Pérez, 2000) y se emplea como **herramienta de aprendizaje complementaria** a las clases impartidas por profesores humanos en euskaltegis (González, Pérez et al., 1999). Salió al mercado en junio de 1999 con 150 puestos contratados distribuidos fundamentalmente en ayuntamientos y euskaltegis. El primer año de comercialización el número de usuarios de Hezinet fue entorno a los 2000 (Sanz-Lumbier, Gutiérrez et al., 2002).

A continuación, a la herramienta de aprendizaje del euskera se le añadieron una biblioteca digital multimedia *CiberBiblio (López-Cuadrado, Villamañe et al., 2001)* y el servidor web de revistas y periódicos digitales *Kiosk@* (Armendáriz, Izquierdo et al., 2002), formando la tripleta educativa que se ha dado a conocer con el nombre de hiperentorno educativo *HEUSKLEARNING* (Armendáriz, López-Cuadrado et al., 2004). El nuevo entorno para el fomento del **autoaprendizaje** del euskera se implantó en recintos denominados **zibergela**k (ciberaulas en castellano). La implantación se hizo antes del verano de 2000 en 21 centros públicos del País Vasco y 14 casas vascas de Argentina y Uruguay, dando servicio a unos 5000 estudiantes (Villamañe, Gutiérrez et al., 2001).

La tercera y más reciente versión del sistema se creó en 2002 bajo el nombre de **Boga.** Esta variante de Hezinet también trata de potenciar el **autoaprendizaje**, sólo que, a diferencia de zibergela, éste no se efectúa necesariamente en un centro físico concreto, sino que se puede realizar **a través de Internet** y esta disponible ininterrumpidamente. Tan sólo se necesita disponer de ordenador y una conexión a la web y pagar la licencia para tener acceso al servidor central. Está versión está instalada no sólo en los *euskaltegis* del País Vasco sino también en centros distribuidos a lo largo de más de 10 países (http://www.boga.habe.org).

Pese a presentar tres modos o ámbitos de funcionamiento diferentes, el núcleo operativo del SHA es el mismo para Hezinet, Zibergela y Boga. Dicho de otro modo, aunque la interfaz de usuario o el modo de interacción difiere en cada caso, la implementación software subyacente no varía, de ahí que **a partir de ahora**, y en esta memoria, **los tres sistemas se englobarán bajo el nombre genérico de Hezinet**.

1.3. Organización y guía de lectura de la memoria

Esta memoria está organizada en cinco partes, cada una compuesta por varios capítulos, tal y como se detalla a continuación:

- La **primera parte** introduce el trabajo de la tesis y presenta el contexto de investigación. Se introduce además el sistema de e-learning de partida: *Hezinet*.
- La **segunda parte** es esencialmente teórica y presenta las *técnicas que se utilizan para la evaluación de sistemas*. Las bases de la experimentación en la ingeniería del software para el desarrollo de los experimentos controlados que generarán sendas calibraciones de ítems. Así como un estado del arte de sistemas de evaluación que utilizan ítems calibrados por expertos y por la TRI para su funcionamiento.
- La **parte tercera** presenta *las calibraciones de los ítems*, la *comparación* de los mismos y propone *procesos de negocio de calibración* aplicable para los casos de la TRI y juicio de expertos. En concreto, el

capítulo 6 describe el experimento desarrollado para construir una calibración de los ítems empleando las valoraciones otorgadas por los expertos. El capítulo 7 recoge el experimento correspondiente para calibrar el mismo banco de ítems según el modelo logístico de 3 parámetros de la TRI. El capítulo 8 contiene un análisis multicriterio de los resultados de los experimentos desarrollados: por un lado, atendiendo a los valores de dificultad estimados, y por otro, atendiendo a los costes asociados a su producción. Finalmente, el capítulo 9 presenta una propuesta de proceso de negocio para la realización de cada una de las calibraciones.

- La **cuarta parte** presenta las *conclusiones* y *aportaciones más relevantes* obtenidas como consecuencia de la realización de este trabajo de investigación, identifica las principales líneas de trabajo a seguir y concluye con las publicaciones generadas.
- Finalmente, la **quinta parte** recoge los *anexos* y las *referencias bibliográficas* citadas a lo largo de esta memoria.

Los bloques de capítulos que componen esta memoria se presentan siguiendo una línea secuencial: tras introducir y contextualizar la investigación (parte 1), se asientan los fundamentos teóricos subyacentes (parte 2) para documentar el trabajo realizado (parte 3) y concluir (parte 4). Aunque el texto está redactado para ser leído al completo y con continuidad, los capítulos contienen citas a conceptos presentados en capítulos precedentes, y es posible que, dependiendo del grado de conocimientos del lector, éste quiera obviar parte de la memoria. En concreto, si está familiarizado con las técnicas de evaluación de software, el capítulo 3 puede no ser relevante; el experto en experimentos controlados puede prescindir del capítulo 4; quien tenga conocimientos de psicometría y calibración puede descartar la lectura del capítulo 5; por último, quien solo desee tener un conocimiento superficial de los resultados de este trabajo, puede centrar su lectura en los capítulos 1, 8, 9 y 10.

Capítulo 2
Hezinet

Hezinet (Pérez, 2000) es un sistema hipermedia adaptativo (SHA) multiusuario y multiplataforma para el aprendizaje del euskera. Utiliza la experiencia del funcionamiento de los Sistemas Tutores Inteligentes (STI) para organizar el conocimiento que se quiere que el alumno aprenda. Por otro lado, utiliza la Teoría de Respuesta al Ítem (TRI) para comprobar la adquisición del conocimiento utilizando tests compilados a partir de bancos de ítems. Y finalmente, presenta al alumno los distintos contenidos utilizando un Sistema Hipermedia Adaptativo (SHA) que le permite navegar según su criterio a través de los materiales multimedia que se le ofrecen.

Hezinet se construyó con el objetivo de promocionar el uso de sistemas de aprendizaje multimedia dentro de la Comunidad Autónoma Vasca, siendo el resultado de un proyecto INTEK de transferencia de tecnología a empresas subvencionado por el Gobierno Vasco/Eusko Jaurlaritza. Se convirtió en un sistema de aprendizaje pionero (Sanz-Lumbier, Gutiérrez et al., 2002) por varios motivos: se trata de un sistema *abierto a toda la sociedad*, *con gran presencia de material multimedia* y *de características adaptativas.*

Primeramente, Hezinet fue concebido como un sistema abierto a toda la sociedad, en contraposición a la mayoría de los sistemas educativos que se desarrollan en proyectos de investigación con uso limitado en escuelas y universidades, y que no llegan a comercializarse.

En segundo lugar, Hezinet dispone de una importante carga de material multimedia (concretamente, en 2001, contaba con 10.231 actividades incluyendo 364 vídeos, 72 películas interactivas y más

703 ficheros de audio) y, desde entonces, se está actualizando constantemente.

Finalmente, el sistema tiene comportamiento adaptativo en varios aspectos. Al comenzar, realiza un test de ingreso para colocar al alumno en el nivel de conocimiento adecuado. Luego, durante el desarrollo del curso, ajusta los contenidos que muestra al alumno dependiendo del desarrollo de la adquisición de conocimientos por parte de éste (Villamañe, Gutiérrez et al., 2001), todo ello utilizando bancos de ítems calibrados.

El capítulo está organizado de la siguiente forma. En la sección 2.1 se presenta cómo se funden las ideas de los STI con la hipermedia para conseguir el aprendizaje que persigue Hezinet. Seguidamente, la sección 2.2 comenta la organización pedagógica del dominio pedagógico seguido de una descripción sencilla del funcionamiento del sistema en la sección 2.3 y los ámbitos de uso del sistema que se están utilizando en la actualidad (sección 2.4). A continuación nos centraremos en aquellos elementos que consideramos críticos dentro del sistema, como las diversas alternativas para efectuar la supervisión del proceso de aprendizaje (sección 2.6) o la evaluación de la adquisición de conocimientos del alumno con el sistema (sección 2.7). Posteriormente, en la sección 2.4, se enuncian las evaluaciones que los desarrolladores han planificado y llevado a cabo sobre el sistema y se finaliza con una breve síntesis de los estadios de evolución por los que ha pasado el sistema.

2.1. Los fundamentos de Hezinet

Los **sistemas hipermedia** integran características de sistemas hipertexto y multimedia. Estos sistemas gestionan la información, organizada en una estructura navegable de nodos y enlaces, utilizando medios audiovisuales de diversa naturaleza. Este modo de organizar la información ofrece múltiples ventajas, como la posibilidad de controlar los contenidos que serán visibles en la siguiente fase del aprendizaje. De hecho, uno de los objetivos principales de los sistemas hipermedia educativos es proporcionar un entrono de aprendizaje que facilite la exploración, de ahí que estos entornos sean idóneos para acceder a grandes colecciones de

información representada en un marco de trabajo conceptual compuesto por nodos y enlaces basados en estructuras semánticas (Pérez, Gutiérrez et al., 2001).

Cuando el sistema hipermedia incorpora mecanismos que le permiten ajustarse dinámicamente a las características del usuario, entonces se dice además que es un **Sistema Hipermedia Adaptativo** (SHA).

La aportación más relevante de Hezinet es la integración de características de los Sistemas Tutores Inteligentes (STI) y de los SHA (Gutiérrez, Pérez et al., 1995). La Figura 1 muestra el esquema de la arquitectura interna del sistema, donde se pueden apreciar los módulos característicos de los STI así como la separación entre la zona adaptativa del sistema de la hipermedia. La simbiosis que se produce es muy interesante ya que aprovecha las ventajas de cada una de las partes: las características hipermedia ofrecen la libertad de acceso a los contenidos y las del sistema tutor controlan al usuario y adaptan el comportamiento del entorno de aprendizaje sin intervenir ni dirigir en exceso la instrucción. Precisamente, Hezinet adapta el hiperespacio disponible, esto es, el conjunto de nodos o sesiones de trabajo que son accesibles en un momento dado, dependiendo del conocimiento del alumno, para, a medida que éste vaya aprendiendo, hacer que la accesibilidad se incremente, dándole la oportunidad de alcanzar nuevas informaciones.

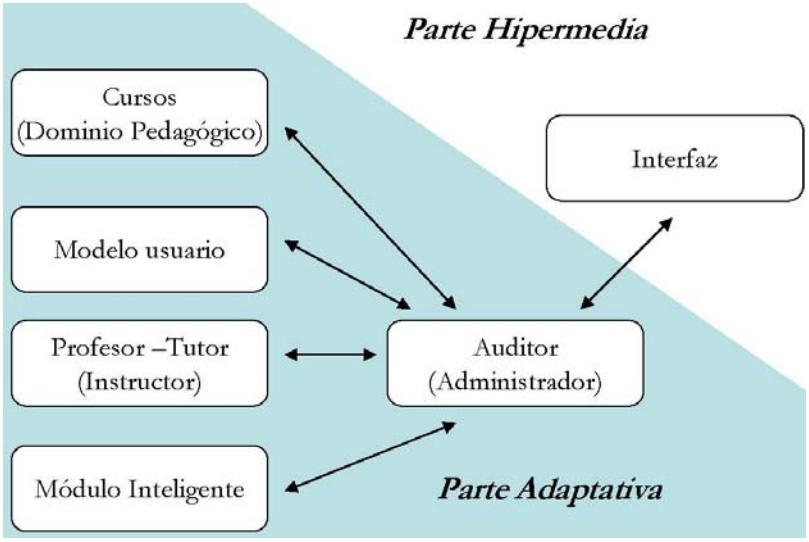


Figura 1.- Arquitectura general de Hezinet

La integración hipermedia-STI igualmente puede ser considerada como una combinación de elementos **constructivos** e **instructivos**

(Gagné y Briggs, 1979; Norman, 1983) que permiten la aparición de distintos paradigmas de aprendizaje, a saber, constructivo, instructivo y **colaborativo** (Pérez, 2000).

Específicamente, en Hezinet las *características constructivas* se concentran principalmente en la cesión del control de la navegación al alumno a través de la red hipermedia interactiva, lo que le permite escoger su propio camino a seguir en su interacción con el entorno. Las características constructivas se complementan, además, con la integración en el sistema de una serie de elementos que soportan, bajo demanda del alumno, actividades de búsqueda de conocimiento. Ejemplo de estos elementos son un diccionario castellano-euskera/euskera-castellano, las hojas de ayuda y un libro electrónico de gramática vasca (Figura 2).

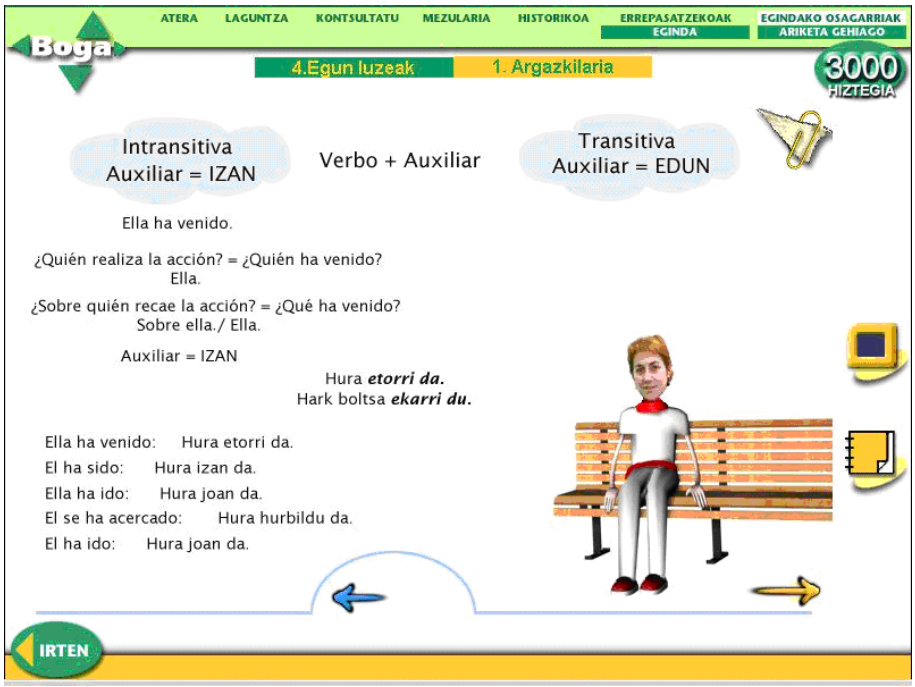


Figura 2.- Libro electrónico de gramática incluido en Hezinet

Las *características instructivas* principales del sistema se encuentran en la descripción del dominio de aprendizaje y su almacenamiento. También existen elementos de supervisión del aprendizaje que, tras acceder a dicha descripción y al histórico del alumno, imponen limitaciones a la navegación en el hiperentorno, bien de manera automática, bien por el modo en que los profesores-tutores emplean el sistema para gestionar la supervisión de la progresión del aprendizaje de los alumnos.

Finalmente, existen también utilidades de consulta con otros alumnos o/y con el profesor-tutor que fomentan el aprendizaje *colaborativo* en Hezinet como son los grupos de discusión en línea (chat) y los grupos de noticias (news).

2.2. La organización pedagógica del dominio

En Hezinet, los conceptos que se desean transmitir se denominan **contenidos**. Un contenido se localiza por dos características asociadas a él: la **destreza lingüística** a desarrollar por el alumno (siendo esta una entre verbos, declinaciones, sintaxis, vocabulario, sufijos, ortografía, conectivas, expresión oral, expresión escrita, comprensión oral y comprensión de textos) y el **nivel de dificultad** de realización. Dentro de los contenidos no todos tienen la misma importancia, por ello se distinguen los denominados *conceptos clave* que son la base de otros conocimientos más complejos.

Las relaciones entre los contenidos se establecen mediante las figuras de grupo y familia. Así, y mientras que los contenidos de un *grupo* están en el mismo **estrato** o nivel de dificultad, una *familia* de contenidos comprende conceptos relacionados con independencia del grado de dificultad de los mismos (Figura 3).

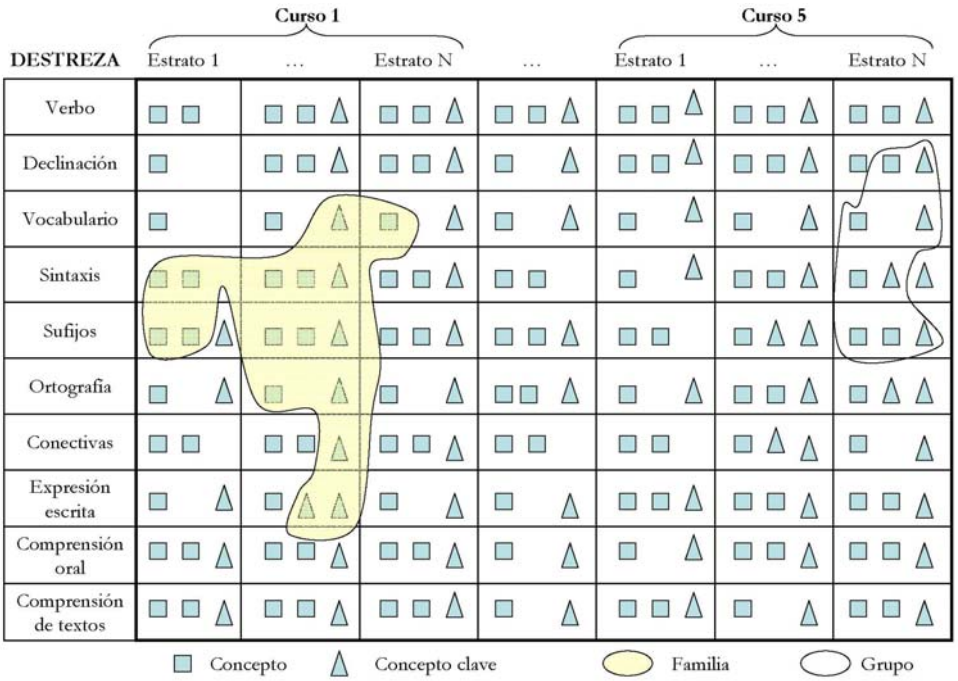


Figura 3.- Ejemplo de estructura de los contenidos del dominio de Hezinet

Hezinet almacena la relación existente entre los contenidos, aunque no anota exactamente en qué consiste dicha ligadura (Pérez, 2000). La Figura 3 muestra una organización hipotética de contenidos del dominio de Hezinet atendiendo a la destreza lingüística y al nivel de dificultad, siendo ambas las únicas características que considera el sistema.

2.3. El funcionamiento de Hezinet

La piedra angular del funcionamiento de Hezinet son las actividades. Una **actividad** es un ejercicio a realizar, basado en algún documento, normalmente, multimedia, cuya respuesta se encuentra en los materiales de consulta que se proporcionan. Su objetivo consiste en suscitar al alumno la necesidad de aprender contenidos que no domina, que se percate de qué contenidos tiene que aprender, que establezca sus técnicas de localización del conocimiento y que sea capaz de aprender de forma autónoma. La resolución de las actividades se hará por transferencia del conocimiento adquirido al problema planteado. De esta manera, la resolución adecuada significará conocimiento que se posee; mientras que lo contrario debiera hacer que el alumno revisara el procedimiento seguido para resolver la actividad puesto que se ha percibido alguna deficiencia en el desarrollo de la misma.

A fin de que las actividades no se hagan repetitivas y permitan al alumno desarrollar su aprendizaje, se han integrado diferentes medios audiovisuales que aportan más frescura en la interacción con el sistema. Además, para que la presentación sea novedosa y evitar desmotivación y aburrimiento, Hezinet utiliza hasta veinte tipos de actividades, que cubren todas las categorías definidas por (Muñiz, 1992): ejercicios de respuesta o elección múltiple, búsqueda de errores en un texto, verdadero/falso, relacionar y/u ordenar elementos (Figura 4), completar espacios en blanco, respuesta corta y ensayo, entre otros (Pérez, 2000).

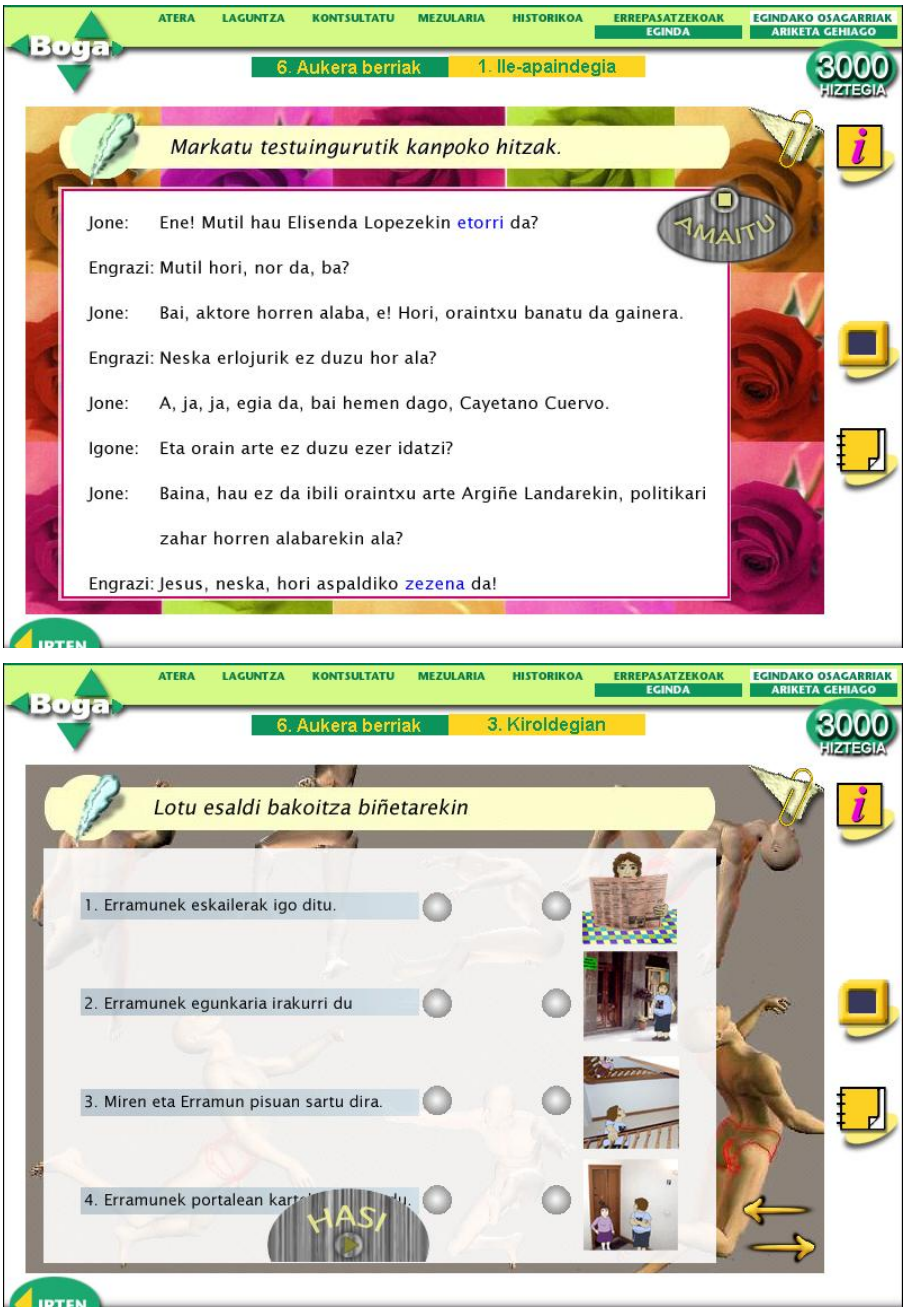


Figura 4.- Ejemplos de actividades de tipo: marcar elementos incorrectos y relacionar con imágenes

A fecha de 2001 Hezinet contaba con más de 10.000 actividades distribuidas en un total de 50 estratos y agrupadas en 5 cursos equivalentes a los necesarios para llegar a obtener el certificado de suficiencia del euskera EGA (*Euskararen Gaitasun Agiria*). Este certificado es el equivalente al *First Certificate* de inglés que emite la Universidad de Cambridge, pero aplicado al euskera. No obstante, indicar que está construido de manera que se podría modificar en cualquier momento el número de cursos o estratos en el dominio y el sistema seguiría funcionando sin problemas.

En la práctica, una actividad puede suponer el trabajo con varios contenidos; aunque en Hezinet se considera que cada una de ellas desarrolla solamente la destreza de uno. Por otro lado, las actividades no se presentan al alumno de manera desordenada, sino que se encuentran agrupadas para trabajar repetidamente los mismos contenidos en otras estructuras que las engloban. La organización de presentación de las actividades se hace a través de sesiones que trabajarán ciertos conceptos contenidos dentro del dominio pedagógico y sobre los que se evaluará si se ha aprendido o no.

Las **sesiones**, con una duración estimada de al menos una hora de trabajo, son los nodos del hiperespacio de Hezinet y están compuestas por actividades que a su vez se han desarrollado como elementos motivadores de la navegación por el hiperespacio (Pérez, Gabiola et al., 1999; Pérez, Gutiérrez et al., 1995b). Así, la navegación simplemente consiste en saltar de sesión a sesión. No obstante, entre sesión y sesión el hiperespacio cambia adaptándose al progreso efectivo del aprendizaje del alumno: si no han aparecido nuevas sesiones (el alumno no ha progresado en su aprendizaje), entonces al menos habrá variado el nodo de la sesión recién visitada. En el ejemplo de la Figura 5 las sesiones visitadas aparecen marcadas en verde y se puede observar que no hay un orden preestablecido de visitas.

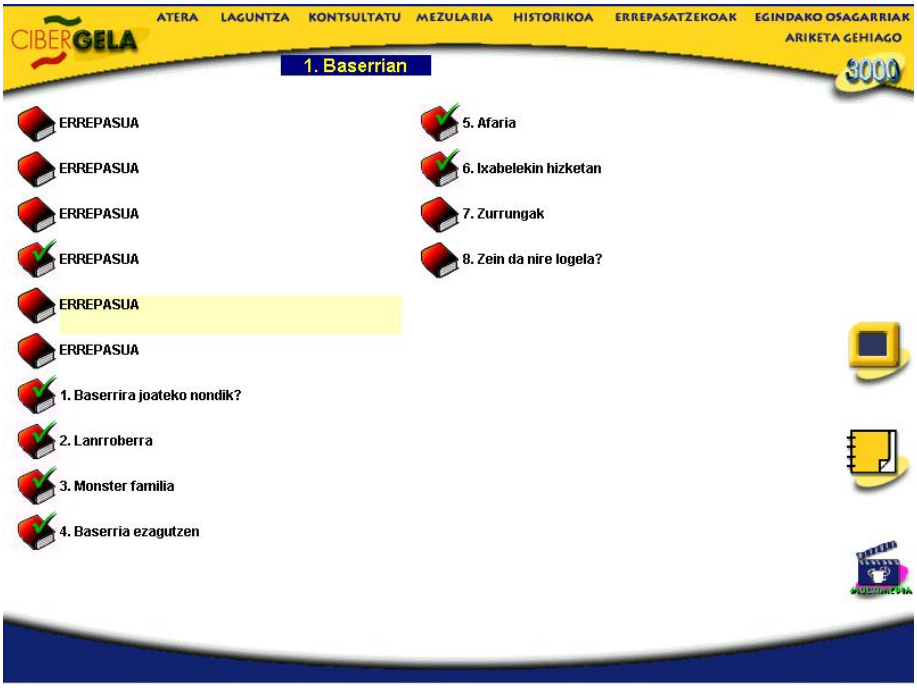


Figura 5.- Las sesiones visitadas se identifican mediante un icono específico

Existen varios tipos de sesiones dependiendo del origen de las mismas.

- **Sesiones predefinidas:** son sesiones creadas por los diseñadores de los contenidos y son comunes para todos los alumnos. Hezinet cuenta con más de 550 sesiones de este tipo.
- **Sesiones de tutor:** estas las crea el profesor para un alumno concreto como consecuencia de algún evento del aprendizaje, como resultado de una consulta, una conversación con el tutor o la corrección de un ejercicio que no se evalúa automáticamente.
- **Sesiones de repaso** ("Errepasua" en la figura): son sesiones generadas por Hezinet de manera autónoma para un alumno y bajo determinadas condiciones.

Consecuentemente, será preciso que las sesiones de Hezinet se puedan realizar independientemente las unas de las otras, lo que se consigue mediante los escenarios pedagógicos (Teusch, Chanier et al., 1996). Las sesiones de Hezinet proponen **escenarios pedagógicos independientes**, a veces incluidos en un contexto más amplio, que colocan al alumno en situación a fin de favorecer el aprendizaje y la práctica de la lengua. La independencia de los escenarios representados en las sesiones es fundamental, ya que evita imponer un recorrido siguiendo una secuencia predeterminada por el hiperespacio.

2.4. Ámbitos de uso de Hezinet

En la actualidad este sistema se usa en tres ámbitos bien diferenciados, cada uno de ellos con una denominación distinta (Hezinet, Zibergela y Boga) aunque con el mismo núcleo operativo.

La versión pionera del sistema se denominó **Hezinet** (Pérez, 2000) y se emplea como **herramienta de aprendizaje complementaria** a las clases impartidas por profesores humanos en euskaltegis (González, Pérez et al., 1999). Salió al mercado en junio de 1999 con 150 puestos contratados distribuidos fundamentalmente en ayuntamientos y euskaltegis. El primer año de comercialización el número de usuarios de Hezinet fue entorno a los 2000 (Sanz-Lumbier, Gutiérrez et al., 2002).

A continuación, a la herramienta de aprendizaje del euskera se le añadieron una biblioteca digital multimedia *CiberBiblio (López-Cuadrado, Villamañe et al., 2001)* y el servidor web de revistas y periódicos digitales *Kiosk@* (Armendáriz, Izquierdo et al., 2002), formando la tripleta educativa que se ha dado a conocer con el nombre de hiperentorno educativo *HEUSKLEARNING* (Armendáriz, López-Cuadrado et al., 2004). El nuevo entorno para el fomento del **autoaprendizaje** del euskera se implantó en recintos denominados **zibergela**k (ciberaulas en castellano). La implantación se hizo antes del verano de 2000 en 21 centros públicos del País Vasco y 14 casas vascas de Argentina y Uruguay, dando servicio a unos 5000 estudiantes (Villamañe, Gutiérrez et al., 2001).

La tercera y más reciente versión del sistema se creó en 2002 bajo el nombre de **Boga.** Esta variante de Hezinet también trata de potenciar el **autoaprendizaje**, sólo que, a diferencia de zibergela, éste no se efectúa necesariamente en un centro físico concreto, sino que se puede realizar **a través de Internet** y está disponible ininterrumpidamente. Tan sólo se necesita disponer de ordenador y una conexión a la web y pagar la licencia para tener acceso al servidor central. Esta versión está instalada no sólo en los *euskaltegis* del País Vasco sino también en centros distribuidos a lo largo de más de 10 países (http://www.ikasten.ikasbil.net/).

Pese a presentar tres modos o ámbitos de funcionamiento diferentes, el núcleo operativo del SHA es el mismo para Hezinet, Zibergela y Boga. Dicho de otro modo, aunque la interfaz de usuario o el modo de interacción difiere en cada caso, la implementación software subyacente no varía, de ahí que a partir de ahora, y en esta memoria, los tres sistemas se englobarán bajo el nombre genérico de Hezinet.

2.5. La supervisión del proceso de aprendizaje

Aunque en Hezinet el propio alumno puede supervisar la evolución de su aprendizaje, no es el único que lo hace. También lo supervisan el sistema, de manera automática, mediante la evaluación de sesiones y estratos; y el profesor-tutor a través de utilidades telemáticas de

comunicación entre profesor y el alumno soportadas por el sistema, o bien a través del trato personal independientemente del sistema. A continuación se comentan en mayor detalle los tres tipos de supervisión que se pueden realizar en Hezinet:

- **Supervisión automática de Hezinet.** Una vez que el alumno realiza todas las actividades incluidas en una sesión, el sistema comprueba la adquisición de conocimientos de éste mediante una evaluación asociada a la sesión y al histórico del alumno. Como resultado de dicha evaluación, el sistema actualiza la lista de contenidos superados, realizados y pendientes para el alumno. Asimismo, cada vez que hay contenidos no superados, el sistema crea automáticamente nuevas sesiones de repaso para fortalecer dichos contenidos.
- **Supervisión del profesor.** El profesor humano puede hacer un seguimiento de la evolución del aprendizaje del alumno consultando el modelo del alumno (Figura 6): las sesiones y actividades realizadas, los tests realizados, las respuestas emitidas junto con los resultados asociados, etc. Puede, igualmente, consultar las actividades que le quedan pendientes. Más aún, el profesor puede gestionar información que la aplicación no puede obtener directamente del alumno, por ejemplo, a través de las herramientas de comunicación incluidas en el sistema o como consecuencia de la corrección de ejercicios que el sistema no puede realizar automáticamente (por ejemplo, corrección de actividades tipo ensayo). Luego, Hezinet permite al profesor hacer labores de supervisión, consultar el modelo del alumno, modificar los datos allí almacenados para registrar la información referente al alumno que el sistema no ha sido capaz de obtener automáticamente, y puede, también, asignar nuevas sesiones a realizar.

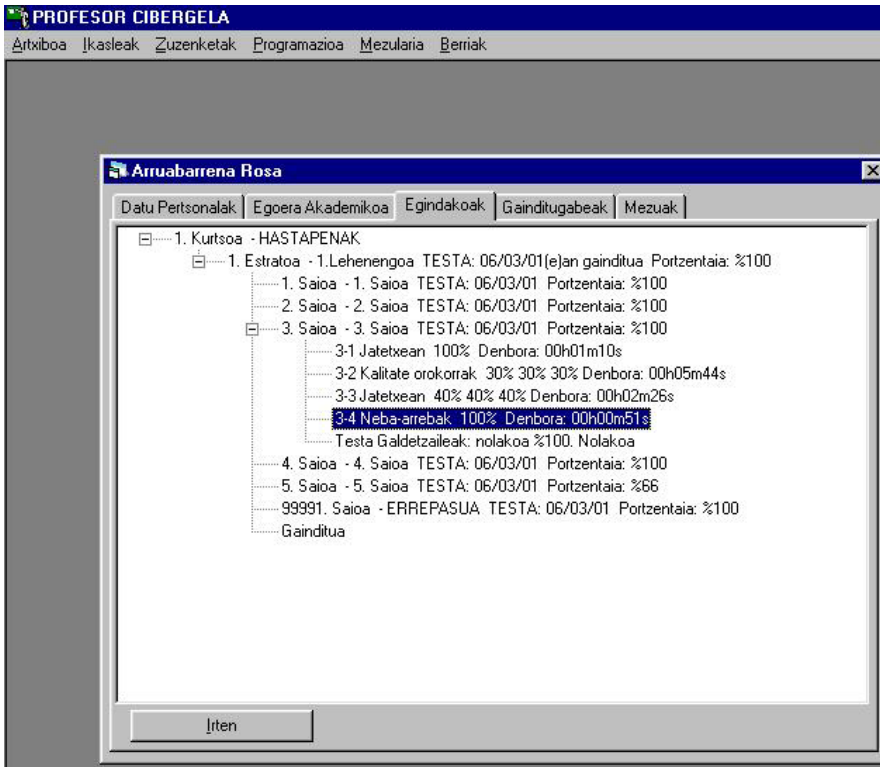


Figura 6.- Hezinet permite al tutor realizar un seguimiento de los progresos del alumno

- **Supervisión del propio alumno.** En Hezinet, el alumno sabe que en cada sesión se trabajan una serie de contenidos y que cada actividad trabaja uno concreto. Sin embargo, puesto que trabajar no es sinónimo de aprender, el alumno necesita de elementos (constructivos) que le permitan autocontrolarse y asegurarse de qué ha aprendido. Por ese motivo, el alumno de Hezinet puede obtener un informe de progreso histórico con el sistema, que da una visión del trabajo realizado en comparación con la totalidad de contenidos, en base al tiempo invertido, el número de actividades realizadas, en comparación con el total de las mismas, y el número total de contenidos a aprender. El informe recoge cuántos de los contenidos se han evaluado y se clasifican según se hayan superado o suspendido. Si durante la realización de una actividad el alumno se percata de que aún no tiene suficiente destreza con el contenido planteado, éste tiene la opción de solicitar al sistema que le proporcione más ejercicios sobre dicho tema. Así mismo, existe otra opción que ofrece también actividades relacionadas con los contenidos tratados. En este caso, las actividades pertenecen a la misma *familia*, por lo que el rango de niveles de dificultad de las mismas puede variar. Estas

dos opciones promueven otra manera constructivista de aprender: la resolución mediante ejemplos.

2.6. La evaluación del alumno

Dado que, por un lado, Hezinet es un sistema destinado al aprendizaje y, por otro lado, uno de los parámetros a los que se debe adaptar el sistema es el conocimiento de los contenidos que se están visionando, se hace necesario hallar un método por el cual obtener datos acerca del proceso de adquisición del conocimiento que se está produciendo. Aunque algunos sistemas, como Elm-Art (Schwarz, Brusilovsky et al., 1996), se decantan por centrarse en el recorrido efectuado por el alumno, en Hezinet se considera que no es acertado suponer que un alumno conoce un determinado concepto tan sólo por el hecho de haber visitado un nodo en el que se habla de él. Se hace necesario un método que certifique de una manera fehaciente que se conocen los conceptos y para ello en Hezinet se utiliza el método habitual dentro del proceso de enseñanza tradicional: el examen. Ahora bien, a diferencia de otros sistemas de enseñanza que utilizan exámenes, las características inherentes de los sistemas hipermedia hacen que no se pueda presuponer a priori el recorrido del alumno a través del hiperespacio a no ser que se limiten las posibilidades de navegación.

Por ello, Hezinet permite compilar los tests de evaluación durante su ejecución mediante un módulo que funciona bajo demanda al que se le puede indicar cuáles son los contenidos sobre los que debe evaluar en un momento dado (Pérez, Gutiérrez et al., 1995a). Este módulo es el encargado de compilar, administrar y corregir los **tests de sesión**, **de estrato** y **de curso**, que evalúan el conocimiento adquirido por los alumnos durante su interacción con el sistema, y presentan ítems sobre los conceptos que se están trabajando, respectivamente, en la sesión, estrato o curso actual. Los ítems que forman parte de estas evaluaciones tienen características similares a las de las actividades utilizadas durante el aprendizaje, y se seleccionan según la correspondencia con los contenidos de las interacciones con el sistema y dependiendo si se han utilizado previamente. Si el alumno contesta correctamente un 60% de las respuestas se considera que éste ha superado los contenidos

trabajados; en caso contrario, se le indica los contenidos que no ha superado y que, por tanto, deberá demostrar posteriormente que los domina.

Hezinet también compila, administra y corrige los **tests de ingreso**, un cuarto tipo de test con una finalidad totalmente diferente a los anteriores, ya que el objetivo de esta prueba es determinar el nivel en el que va a comenzar el alumno a interactuar con Hezinet. Mientras que los tres primeros tests están orientados a evaluar un punto concreto del aprendizaje, el test de ingreso al sistema incluye preguntas sobre los conceptos clave de cada uno de los cursos del programa educativo. Sin embargo, el test de ingreso requiere un tratamiento especial ya que el número de conceptos clave puede llegar a ser excesivo, pues cubre todos los conceptos clave de cada uno de los cursos que se ofrecen, y se hace necesario utilizar técnicas de evaluación adaptativa por ordenador para reducir la longitud del test sin dañar la fiabilidad de sus resultados (Hambleton, Swaminathan et al., 1991). Además, todos los ítems o preguntas utilizados en este tipo de evaluación son de respuesta múltiple, pues de otro modo resultaría más difícil garantizar resultados fiables desde el punto de vista psicométrico.

Para poder construir un test es preciso disponer de una colección de actividades habitualmente almacenada en una base de datos, y que en este contexto se denomina **banco de ítems**, de las que se seleccionaran los ejercicios que conformarán el test. El método de compilación de los ejercicios que componen un test de Hezinet se basa en la TRI (Lord, 1952; Lord, 1980). El compilador puede utilizar de 0 a 3 de los parámetros contemplados para construir un test (dificultad, discriminación y pseudo-acierto). Tan sólo depende de los datos que tengan almacenados los ítems. En principio se parte de 0 y, a medida que se recojan datos, se estimarán los parámetros y se comenzará a utilizar más parámetros de los disponibles. En Hezinet se ha construido un módulo *compilador de test* al que se le indica cuál es el banco de ítems con el que cuenta para realizar la evaluación. Además, se utiliza un dato extra; como se realizan múltiples evaluaciones en el tiempo, es crucial que el alumno no tenga que resolver dos veces el mismo ejercicio, ya que los resultados de la evaluación se falsearían porque tras realizar cada ejercicio, el alumno tiene la posibilidad de ver qué opciones fueron

consideradas incorrectas y cuál era la respuesta correcta. Por esta razón, se ha incluido en el algoritmo de compilación la directriz de evitar la repetición del uso del mismo ítem (en la medida de lo posible) con un alumno, o distanciar, al menos, su utilización al máximo en el tiempo (Pérez, Gutiérrez et al., 1995b). De esta manera el alumno no puede memorizar las respuestas a las preguntas como método para superar un test. En concreto en el 2001, Hezinet contaba con un banco de más de 600 ítems con actividades sobre los contenidos de los que constaba el sistema para compilar ejercicios de test.

2.7. Evaluación del sistema

Así, y desde su concepción, la aplicación Hezinet ha soportado varias evaluaciones. Concretamente, como consecuencia de una primera evaluación formativa la interfaz de Hezinet se modificó substancialmente y se le añadió alguna funcionalidad más al sistema antes de su lanzamiento al mercado (Pérez, López et al., 2000). Tanto su rápida expansión a nivel mundial tras su comercialización como su integración con otro tipo de aplicaciones relacionadas con la cultura vasca (Armendáriz, López-Cuadrado et al., 2004) avalan de manera informal la validez del sistema.

En esta situación, se realizó un estudio con vistas a identificar posibles áreas de mejora, y se elaboró un plan de evaluación sumativa que se recogió en (Arruabarrena, Pérez et al., 2001; Arruabarrena, Pérez et al., 2002). Específicamente, las evaluaciones sumativas planificadas fueron: (1) medir el nivel de amigabilidad que presentaba la interfaz, (2) concretar la incorporación o mejora de nuevas prestaciones del sistema, (3) evaluar el impacto afectivo de Hezinet sobre los distintos tipos de usuarios del sistema, (4) medir la efectividad del sistema en el progreso del aprendizaje de los distintos tipos de alumnos que trabajaron con y sin el sistema en los euskaltegis y (5) valorar la integración de la herramienta dentro del proceso de aprendizaje del euskera desarrollado en los euskaltegis.

Relacionado con el plan de evaluación de Hezinet propuesto, en (Villamañe, Gutiérrez et al., 2001) se presentó una comparación entre los resultados obtenidos en exámenes oficiales por alumnos de

euskaltegis que habían empleado la herramienta como apoyo al aprendizaje del euskera y los que no. Desde el punto de vista de efectividad del sistema, que los usuarios de Hezinet obtuvieron mejores resultados; y, desde el punto de vista afectivo, consideraron muy positiva la experiencia con el sistema. Así mismo, en este estudio se observó que el curso uno de Hezinet, el de nivel más básico, era demasiado elevado para los alumnos que desconocían totalmente el idioma. Boga, la versión online siguiente a Hezinet, tiene incorporado el curso cero y presenta una nueva interfaz acorde con las últimas tendencias gráficas.

PARTE 2
Fundamentos

*La **Parte 2** está dedicada a presentar los fundamentos teóricos en los que se sustenta la propuesta de procesos de negocio para realizar calibraciones de ítems válidas y eficientes. Hoy por hoy los dos procesos de calibración más extendidos son la opción que emplea valoraciones de expertos y la que aplica procesos estadísticos psicométricos para realizar una estimación de los valores de los parámetros de los ítems.*

*El **capítulo 3** subraya la necesidad de mejorar el proceso de desarrollo de software, ya que incide directamente en la calidad del producto final y en la reducción del esfuerzo de desarrollo. Para mejorar el software o su desarrollo es preciso realizar las tareas de análisis, identificación, verificación y validación del mismo con rigor científico y sistemáticamente. Así, y tras establecer los objetivos de la mejora y el tipo de evaluación a realizar, se podrán emplear y combinar varias estrategias de validación junto con técnicas de evaluación y captación de información como las presentadas en este capítulo.*

*El **capítulo 4** expone los principios fundamentales de la experimentación en el área de la ingeniería del software. También justifica la necesidad de la experimentación como mecanismo para validar nuevos procesos, métodos o herramientas. Se enuncian además los procedimientos de normalización de valores más extendidos y las pruebas estadísticas para realizar comparaciones de resultados más frecuentemente empleadas en esta disciplina.*

*El **capítulo 5** está íntegramente dedicado a los bancos de ítems calibrados. Se muestran sus orígenes, utilidad, áreas de aplicación y algunos sistemas de aprendizaje que los incluyen. Posteriormente se abordan las dos alternativas de calibración más extendidas: la calibración de ítems empleando aportaciones subjetivas de expertos y la estadística en el marco de la TRI.*

Capítulo 3
Evaluación y mejora en la Ingeniería del Software

La mejora del desarrollo del software incide directamente en la calidad del producto final y en la reducción del esfuerzo de desarrollo; y es por ello que el interés suscitado en la década de los 70, relativo a la mejora del desarrollo de productos software en el ámbito industrial, persiste en la actualidad, alcanzando también el ámbito académico y el de la investigación. La complejidad intrínseca del software causa que las evaluaciones sean determinantes para aumentar la calidad del producto final. No obstante, el desarrollo del artefacto no termina en este punto. Los incesantes avances tecnológicos hacen imprescindible ejecutar mejoras continuas sobre los mismos, siendo requisito que el proceso original esté bien definido y documentado. Para mejorar el producto software es preciso realizar tareas de análisis, identificación, verificación y validación del mismo con rigor científico y sistemáticamente. Para ello, y tras establecer los objetivos de la mejora y el tipo de evaluación a realizar, se podrán emplear y combinar varias estrategias de validación junto con técnicas de evaluación y captación de información como las propuestas en los siguientes apartados.

En este capítulo se exponen distintos enfoques, métodos y herramientas utilizados para evaluar sistemas software. Los resultados del estudio se han obtenido principalmente de la literatura del área de sistemas de e-learning, todo ello con vistas a mejorar el producto software Hezinet. Sin embargo, la magnitud de la cuestión es tal, que en este capítulo únicamente se dan unas pinceladas centradas en sistemas EAO, puesto que como efecto lateral del presente trabajo se pretende mejorar Hezinet ampliando sus prestaciones de forma eficaz y eficiente.

3.1. Estrategias de validación

Validar un programa consiste en demostrar que realiza las tareas para las que ha sido creado. Esta demostración se puede realizar de varias formas que se denominarán *estrategias de validación*. Igualmente, una estrategia puede llevarse a cabo de varias maneras distintas, a las que se les llamará *técnicas de evaluación*, y que se verán en el siguiente apartado. Los autores (Mark y Greer, 1993) han considerado las siguientes estrategias de validación:

- **Demostración formal de corrección.** Consiste en emplear técnicas formales de corrección y verificación de programas. Para ello es necesario que se especifiquen de manera formal los requisitos del programa, elemento del que suelen adolecer los sistemas basados en inteligencia artificial, en parte porque muchos no están completamente especificados.
- **Validación basada en el criterio.** Se trata de comprobar si el sistema no presenta grandes insuficiencias con respecto a los objetivos (Patridge, 1986) y para ello se realizan una serie de pruebas con él. Si las supera, se puede asegurar que el sistema es exitoso. Si falla alguna de ellas, el sistema no lo es tanto. Se debe definir un criterio que determine la validez del sistema frente a las pruebas. Por ejemplo, un sistema se considera válido si supera todas las pruebas, o si falla como mucho dos de ellas, etc. Esta estrategia se utiliza con frecuencia en la ingeniería del software para validar los sistemas de información. Es una tarea complicada, normalmente realizada por expertos conocedores del entorno donde se va a utilizar el sistema.
- **Revisión de expertos.** En este caso existe una persona o un conjunto de ellas que supervisa el sistema completo, partes del mismo, su comportamiento, etc. para localizar deficiencias y mejorar la aplicación. Los expertos pueden efectuar su evaluación centrándose en uno o varios módulos del sistema, en la arquitectura del mismo, las tareas que se pueden realizar con él, etc. Los expertos ofrecen información basada en su experiencia, lo cual redunda en credibilidad a la hora de validar un sistema (Shneiderman, 1998). El resultado es un *informe de asesoría* en el que un experto emite su opinión acerca del sistema que ha evaluado. A veces, diferentes expertos pueden dar lugar a un *panel de expertos* donde expondrán sus diversas opiniones sobre el sistema e incluso consensuarán algunas de ellas. Hay estudios,

como (Davidove y Reiser, 1991), que confirman que una revisión de los materiales didácticos vía expertos puede mejorar la eficiencia del aprendizaje.

- **Validación basada en la certificación.** Se supone que debe haber una autoridad certificadora que asegura que el programa se comporta con respecto a ciertos estándares (J.C.S.E.E., 1994), determinando si el sistema es o no competente. Esta estrategia se basa en técnicas de calidad. No es adecuado utilizarla mientras no se tengan estándares para juzgar programas, criterios para evaluar sistemas y sus módulos, y mientras no se sepa identificar programas educativos eficaces de forma precisa (Mark y Greer, 1993). En estos objetivos están trabajando actualmente diferentes organismos internacionales, entre los que cabe mencionar el proyecto SQuaRE (Software product Quality Requirements and Evaluation) de la CERT-Carnegie Mellon University, y que en el ámbito informático va a dar lugar a una serie de estándares, concretamente la familia ISO/IEC 2500n que ofrecerá un modelo general de referencia sobre requisitos y evaluación de la calidad del software. Sin embargo, hasta que las normas de esta serie sean publicadas, las familias ISO/IEC 9126 y la ISO/IEC 14598 son los estándares a utilizar para definir un modelo de calidad de producto y su evaluación (Plaza, Marcuello et al., 2007).
- **Pruebas empíricas.** Mediante esta técnica se examinan las relaciones entre los usos de las aplicaciones y los resultados en los usuarios. Esta estrategia surge cuando es complicado comprobar la correcta adecuación a los requisitos del sistema, por ejemplo, en sistemas educativos, entre las intervenciones instructivas y los resultados de enseñanza en los estudiantes. Hay diversas formas de llevar las pruebas empíricas a cabo, como son el *diseño basado en un único grupo,* el *diseño basado en un grupo de control* y el *diseño cuasi-experimental*. La revista Empirical Software Engineering, de la Kluwer Academia Press, es una publicación periódica donde se pueden encontrar diversos tipos de estudios empíricos. Relacionado con esta estrategia de validación está el diseño experimental en el ámbito de la ingeniería del software, asunto al que está dedicado el Capítulo 4 del presente trabajo.

3.2. Tipos de evaluaciones de sistemas

La evaluación de un sistema educativo es un proceso de recogida de datos con los cuales se realiza una tasación de la valía de la instrucción, y de sus puntos débiles y sobresalientes (Tessmer, 1993). Existen varias aproximaciones en la literatura sobre el tema. (Mark y Greer, 1993) y (Draper, Brown et al., 1996) hacen una recopilación de algunas de ellas. Otros autores (Cawsey, Jones et al., 2000; Welch y Brownell, 2000) realizan evaluaciones empíricas. También hay autores, como (Eklund and Brusilovsky 1998; Calvi 2000) que proponen algún elemento novedoso, sobre todo en los sistemas adaptativos. Específicamente, la más utilizada ha sido "con y sin", en la cual una instancia adaptativa del sistema es comparada frente a otra no adaptativa. Ejemplos de este tipo de evaluación se hallan en (Boyle and Encarnación 1994; Meyer 1994; Weber and Specht 1997; Brusilovsky and Pesin 1998). No obstante, el principal inconveniente que se encuentra a esta teoría es que la adaptación no es una característica que se pueda eliminar fácilmente, sino que es una característica inherente (Höök, 2000). Más recientemente, (Karagiannidis and Sampson 2000; Brusilovsky, Karagiannidis et al. 2001; Paramythis, Totter et al. 2001; Weibelzahl 2001) han propuestos diversas metodologías para efectuar evaluaciones de sistemas educativos adaptativos. Las cuatro metodologías propuestas por los respectivos autores coinciden en la realización de la evaluación por capas o módulos, si bien cada uno propone un número desigual de las mismas y con distinta granularidad. Hay que puntualizar que a día de hoy ninguna de las propuestas se ha impuesto sobre las otras.

A partir de las referencias indicadas así como de manuales que indican cómo realizar evaluaciones de calidad de currícula educativos (Harvey, 1998), se ha construido una clasificación de los distintos tipos de evaluación. Se han distinguido dos clasificaciones de los sistemas de educación: una atendiendo al objetivo que persigue la evaluación (*paradigma de evaluación*) y otra según qué elementos del sistema de educación se evalúan (*objetos evaluados*). Ambas clasificaciones se detallan en los siguientes apartados.

3.2.1. Paradigmas de evaluación

Un paradigma de evaluación es una manera de conducir una evaluación en la que los objetivos ya están definidos explícita o implícitamente. Los paradigmas se distinguen unos de otros por los objetivos en los cuales se centra la evaluación. En la bibliografía se distinguen varios paradigmas que se han resumido en la lista siguiente (Scriven 1991; Mark and Greer 1993; Shute and Regian 1993; Draper, Brown et al. 1996):

- **Evaluación formativa**, de seguimiento o de proceso. Evalúa el sistema durante el proceso de construcción del mismo. Se trata de identificar los puntos débiles para hacer las modificaciones oportunas antes de tener el producto finalizado; luego, su objetivo se dirige a la mejora y optimización del programa. Los diseñadores instruccionales recomiendan que la evaluación formativa debería comenzar en las primeras fases del desarrollo del software, antes de que se hayan invertido sustanciales cantidades de tiempo y dinero (Gagné, Briggs et al., 1988). Con frecuencia se ve como parte del ciclo de vida del software: diseño, implementación y evaluación formativa (McGraw y Harbison-Briggs, 1989). Se podría expresar de manera metafórica (Harvey, 1998) como un cocinero que prueba la comida durante el proceso de elaboración de la misma, para comprobar la calidad del producto que está elaborando. (Calvi, 2000) presenta una evaluación de este tipo.
- **Evaluación sumativa**, de resultados o de impactos. Mide de forma puntual la efectividad del sistema. Los parámetros de evaluación son costes, beneficios y objetivos prefijados del sistema. Se suele llevar a cabo cuando se ha desarrollado el sistema y ante los resultados que ofrece. Siguiendo con la metáfora del punto anterior, este tipo de evaluación es la que, una vez el cocinero ha terminado de elaborar el manjar, los comensales lo prueban, evaluando la exquisitez de la comida que se les ha preparado, valorando, además, el precio pagado, las expectativas satisfechas, etcétera. Evaluaciones de este tipo se pueden encontrar en (Cawsey, Jones et al., 2000) y (O'Hanlon, 1999).
- **Evaluación iluminativa.** Este tipo de evaluación ayuda a identificar los factores y resultados inesperados de una situación concreta. Está inspirada en los métodos e investigaciones etnográficas (Parlett y Haminlton, 1987). Siguiendo con la metáfora del cocinero, este enfoque la evaluaría en un restaurante determinado, con su clientela

habitual e intentaría descubrir factores que son importantes para los clientes de este restaurante, pero que a los cocineros se les había pasado por alto. En (Ewing, 2000) el autor describe la importancia de este tipo de estudios y cómo lo emplea en su trabajo.

- **Evaluación de integración.** Mide el grado de integración que hay entre un software educativo y los demás materiales que se utilizan en el entorno. Esto es, no sólo se mide el sistema, sino también los métodos de aprendizaje que se usan paralelamente (Draper, Brown et al., 1996). Metafóricamente, diríamos que no sólo se trata de catar la comida y valorarla, sino también de evaluar el servicio de camareros, la limpieza de la vajilla, el local donde uno se encuentra degustando la comida, el lugar donde está el local, etcétera (Reid y Arends, 1998) y (Taylor, Woodman et al., 2000) presentan evaluaciones de este tipo en sus respectivos sistemas.

3.2.2. Objeto evaluado

(Littman and Soloway 1988; Mark and Greer 1993; Murray 1993) organizan las evaluaciones de los STIs dependiendo de los elementos del sistema sobre los que se va a centrar la evaluación. Se distinguen tres tipos de evaluaciones:

- **Evaluación interna.** Se evalúa el sistema en concreto. Pueden evaluarse, por ejemplo, cada uno de los componentes del sistema, la arquitectura, los procesos intermedios, los comportamientos y relaciones existentes entre los anteriores. Normalmente, este tipo de valoraciones se encuentra asociada a la evaluación formativa porque la valoración de los distintos módulos se realiza durante el diseño, desarrollo e implementación del sistema. (Or-Bach y Bar-On, 1993) ofrece este tipo de evaluación.
- **Evaluación externa.** Se valoran elementos externos al sistema de aprendizaje, como, por ejemplo los logros alcanzados por el estudiante o el impacto afectivo del sistema. Se mide la adquisición de conocimiento, su comprensión, actuación y transferencia, emociones que han podido impresionar al alumno, sus actitudes... Ejemplos de este tipo de evaluación se pueden encontrar en (O'Hanlon, 1999) y (Reid y Arends, 1998). Para llevarla a cabo, es necesario que el sistema se haya desarrollado completamente, por lo que se estará hablando de una evaluación sumativa.

- **Evaluación global.** En este caso se valora tanto los componentes del sistema como el impacto del mismo en los usuarios. Equivale a hacer una evaluación interna y externa del sistema. (Welch y Brownell, 2000) hacen una evaluación de este tipo.

3.3. Técnicas de evaluación de sistemas

Las técnicas de evaluación son el método concreto con el que se lleva a cabo la estrategia de validación del sistema. También pueden considerarse como herramientas para capturar información y/o conocimiento. Se pueden combinar varias técnicas o herramientas, por ejemplo, un experto puede realizar la comparación de dos sistemas y hacer una prueba piloto. En este apartado se muestran técnicas a utilizar durante el periodo de validación de un sistema; la mayoría se han recogido de (Murray 1993; Shute and Regian 1993; Nielsen and Mack 1994; Shneiderman 1998) y (Harvey, 1998).

3.3.1. Comparación

Consiste en cotejar las características de un sistema con las de otro. El resultado sirve para resaltar las similitudes y las diferencias de comportamiento o de diseño entre ambos. Existen varios métodos, como:

- **Estándar de oro.** La comparación se hace con respecto a otro sistema de reconocido éxito o **prestigio** (el estándar).
- **Corroboración teórica.** Demuestra de manera teórica que un nuevo sistema se puede utilizar usando una nueva aproximación.
- **Corroboración empírica.** Se compara el sistema con una variante del mismo obtenida por ablación (eliminación), extensión (adición), sustitución o modificación de algún módulo o elemento (Calvi, 2000).
- **Duplicación.** La aplicación emula a otro sistema anterior hasta llegar a un nivel adecuado de concordancia, y a partir de ese punto, el sistema puede extenderse o modificarse para intentar mejorar el anterior. Este método es el utilizado por muchos programas de

software que ofrecen compatibilidad con sistemas anteriores, pero siempre añadiendo nuevas características en las nuevas versiones.

- **Test de Turing.** Compara el comportamiento humano y el de la computadora. Se considera que el sistema es válido cuando su comportamiento es indistinguible del humano o superior (Parry y Hofmeister, 1986). Una posible forma de implementar el test de Turing es utilizando la *caja del mago de Oz*, que consiste en usar una persona para simular el comportamiento del sistema propuesto (Murray, 1993). Como ejemplo de utilización de esta herramienta están (Virvou y du Boulay, 1999) y (Green y Carberry, 1999). Mientras que en la primera referencia se compara el sistema con un experto humano en reconocimiento de planes, en la segunda se comparan las alternativas de respuestas del sistema con las alternativas humanas.
- **Análisis de sensibilidad.** Evalúa los cambios de comportamiento del sistema frente a cambios en las entradas. Es decir, se valora la sensibilidad del sistema a cambios en el entorno (Or-Bach y Bar-On, 1993). Cuanto menores sean las diferencias entre los distintos grupos de datos de la entrada que provocan una modificación en el comportamiento, más sensible será éste.
- **Benchmarking.** Consiste en utilizar un banco de pruebas estándar sobre el que se van a evaluar características concretas del sistema. Esta manera de establecer comparaciones supone que los sistemas comparados tienen muchos procesos comunes de modo que se pueda someter a todos ellos a las distintas pruebas (Shute y Regian, 1993).

3.3.2. Contacto con usuarios

El contacto con el usuario puede ofrecer mucha información. Se trata de obtener datos acerca de la manera de trabajar que tiene el usuario con el sistema. Los datos obtenidos sobre un sistema pueden ser tanto cuantitativos como cualitativos (opiniones). Las técnicas disponibles son:

- **Entrevistas, encuestas y cuestionarios.** Una o varias personas (los administradores, entrevistadores encuestadores) recogen una serie de datos acerca de las ideas de otros individuos (los administrados) sobre un cierto tema o aspecto de interés. La diferencia entre las tres técnicas radica en la persona de la que parte la iniciativa de hacer las preguntas y el método de selección de los individuos que las responden. En las entrevistas y las encuestas la iniciativa de conseguir

las respuestas parte del administrador mientras que en los cuestionarios es el administrado el que decide aportar su opinión. El método de selección de los administrados suele variar: desde la decisión del administrador según sus propios criterios en las entrevistas, de manera aleatoria para representar a una población en las encuestas y según decisión del administrado en los cuestionarios.

- **Test.** El administrador establece una serie de preguntas para conocer (a menudo graduando) los conocimientos o aptitudes del administrado. Se pueden realizar de manera independiente o regulada. Un ejemplo típico son los pre-tests junto con los post-tests que permiten comprobar diferencias antes y después de un determinado evento.
- **Grupos de enfoque y grupos nominales.** Consisten en un grupo moderado de personas relacionadas con un cierto evento de interés. En el primer grupo, los miembros disertan. En el segundo, el evaluador recoge de cada participante del grupo una opinión o reflexión escrita con respecto a una afirmación o pregunta formulada por él. A continuación, las expone a todos los participantes y se votan, pudiendo dar alternativa a un intercambio de opiniones a posteriori. Se emplean para generar información nueva, fresca, en el sentido de que no haya sido restringida por formatos de evaluación preconcebidos, sirviendo para identificar necesidades diversas. Son útiles para mejorar los diseños de las encuestas y el conocimiento de las necesidades de los clientes, y para aumentar el manejo de conceptos para planificaciones futuras.
- **Mapas conceptuales.** Son una representación visual de los enlaces o asociaciones ente conceptos o piezas de información distintas, con aplicación en áreas muy diversas. En el área de la enseñanza se pueden emplear bien al final del tutorial (dedicando unos pocos minutos) para consolidar el aprendizaje y comprobar la compresión, bien al inicio de la sesión para situar el entorno a los estudiantes (Harvey, 1998). Las opiniones de estos, a su vez, pueden poner de manifiesto aquellos conceptos o puntos que no están suficientemente clarificados desde el punto de vista del usuario.
- **Think aloud.** El evaluador solicita al usuario que realice una serie de tareas concretas y que, simultáneamente, vaya exponiendo sus pensamientos. El evaluador da soporte a los usuarios, mas nunca toma el control de la situación ni da instrucciones. Su labor se limita a

observar, escuchar y registrar todos los comentarios. Los datos del experimento se pueden recoger mediante protocolos distintos: lápiz y papel, por grabación de audio y/o video, y por registro en ordenador. La prueba se realizará en un entorno informal y tras su finalización, se generará un informe donde quedarán reflejadas todas las experiencias de los participantes.

- **Método DELPHI[1].** El método consiste en interrogar a expertos con la ayuda de cuestionarios, a fin de poner de manifiesto convergencias de opiniones y deducir eventuales consensos. Los expertos no trabajan físicamente juntos, sino que cada uno de ellos opina por escrito en un ambiente de anonimato que facilita su libertad de expresión. Es una metodología de investigación multidisciplinar para la realización de pronósticos y predicciones Está caracterizado como un método para estructurar el proceso de comunicación grupal, de modo que ésta sea efectiva para permitir a un grupo de individuos, como un todo, tratar un problema complejo (Linstone y Turoff, 1975). La capacidad de predicción del método se basa en la utilización sistemática de un juicio intuitivo emitido por un grupo de expertos. El objetivo de los cuestionarios sucesivos es disminuir el espacio intercuartil; esto es, cuánto se desvía la opinión del experto de la opinión del conjunto, precisando la mediana de las respuestas obtenidas. La técnica se ha convertido en una herramienta fundamental en el área de las proyecciones tecnológicas, incluso en el área de la administración clásica y operaciones de investigación. En (Landeta, 2006) se revisa la validez del método avalado por más de un millar de publicaciones desde su creación y presenta tres aplicaciones recientes, de carácter profesional y poco habituales, dentro de los campos en los que se suele emplear el método.
- **Metodología PERT y distribuciones Beta.** Es una metodología de investigación que utiliza tres estimaciones subjetivas (los valores optimista, pesimista y más probable) aportadas habitualmente por expertos, para la realización de pronósticos y predicciones en la

[1] Método DELPHI. Método ideado por el centro de investigación estadounidense Corporación RAND (Research ANd Development) al inicio de la guerra fría, en los años 50, para investigar el impacto de la tecnología en la guerra, siendo sus artífices T.J. Gordon, Olaf Helmer y Norman Dalkey. El nombre del método se inspira en las predicciones del antiguo oráculo de Delphos.

resolución de problemas de valoración económica, tasación e inversión (incluidos planes de pensiones, de inmobiliarias, valoración de los tiempos en la realización de una tarea, etc.). El empleo de la metodología PERT permite a posteriori comparar predicción con resultado real, esto es, permite contrastar la bondad de un experto valorando la acuracidad de sus predicciones (Herrerías, Palacios et al., 1999). En la literatura sobre la metodología PERT han aparecido cuatro subfamilias de distribuciones beta atendiendo al modelo probabilístico subyacente, a saber, la clásica, la de varianza constantes, la mesocúrtica y la Caballer. Hasta ahora, estas cuatro subfamilias han sido utilizadas independientemente, discriminándose su uso sólo en función de las medias y las varianzas obtenidas en cada una de ellas. Así, cuando se desea emplear un criterio de prudencia se considerará la varianza máxima, o ante una posición más arriesgada se preferirá una varianza mínima. Con respecto a la media, interesa que ésta sea la más cercana al centro del intervalo, es decir, el modelo que proporciona un valor esperado más centrado y, por tanto, más moderado en sus estimaciones (García-Pérez, Cruz-Rambaud et al., 2004).

3.3.3. Análisis de datos

El análisis de datos consiste en estudiar de forma sistemática una serie de datos acerca de ciertas características de interés. Los datos pueden tener carácter cuantitativo o cualitativo, e independientemente de dicho carácter y para que se puedan cubrir eficazmente los propósitos y usos de la evaluación, la información deberá ser analizada sistemáticamente y de forma precisa (Frechtling y Sharp, 1997).

Si la información tiene formato cuantitativo, ésta se tratará empleando algún análisis cuantitativo, para organizarla, sintetizarla e interpretarla. Se emplean generalmente técnicas estadísticas y de síntesis apropiadas para que se ajusten los objetivos que quiere valorar la evaluación con la naturaleza de los datos numéricos.

Si la información recopilada tiene formato cualitativo, ésta se tratará empleando técnicas de análisis cualitativo con los mismos objetivos que la cuantitativa, añadiendo, además, la tarea de descifrar la información narrativa o gráfica para que pueda ser interpretada correctamente y ello ayude en los propósitos de la evaluación. Hoy en día, aunque ya hay

procedimientos sistemáticos para reducir los datos, no siempre se puede efectuar aplicando unas reglas pre-especificadas.

Dentro del análisis de datos se encuentra, por ejemplo, el estudio de historiales de usuarios, que dan idea de las preferencias del usuario, de su "modus operandi" con un software y hace que los diseñadores tengan en cuenta aspectos de la interacción que de otro modo serían ignorados. Para ello, los propios sistemas pueden grabar un historial donde se almacena, paso a paso, la interacción del usuario con la aplicación. Normalmente, se evalúan las interacciones con los elementos de la interfaz: clicks del ratón, selecciones de menú, visitas a nodos concretos... Del mismo modo, y como indicador del desarrollo del aprendizaje, se pueden estudiar los niveles de confianza del usuario en algún punto particular de la instrucción.

3.3.4. Pruebas piloto

Las pruebas piloto consisten en estudiar la actuación del sistema con algunos de los que serán los futuros usuarios del mismo. Se emplean para determinar si el sistema funciona realmente como se había previsto y para asegurarse de que las respuestas inesperadas no den pie a reclamaciones formales, pudiendo identificar problemas e intereses de los usuarios. Dependiendo del número de usuarios que se evalúan de manera simultánea se habla de pruebas piloto *uno a uno* (una persona), de *grupos reducidos* (pocas personas) o *pruebas de campo* o *test beta* (muchas personas) (Tessmer, 1993).

- **Prueba piloto uno a uno**. El evaluador observa detalladamente cómo el usuario interacciona con el sistema desarrollado. Cada vez se trabaja con un único usuario, quien hace las veces de crítico y aprendiz. Los investigadores pueden observar las capacidades y habilidades del usuario e identificar falsas expectativas. Igualmente, pueden emplear la técnica para detectar instrucciones, preguntas e informaciones confusas así como características o situaciones inesperadas. Suele emplearse en la fase inicial del desarrollo del software con vista a minimizar desarrollos inadecuados. Según (Tessmer, 1993), la diferencia entre este tipo de prueba piloto y la revisión de expertos radica en el enfoque distinto de las evaluaciones para los cuales se emplean las técnicas. De este modo, las pruebas piloto uno a uno se emplean cuando lo que interesa es evaluar la

claridad del sistema, la facilidad del uso y la eficiencia del aprendizaje del sistema por parte de los usuarios.

- **Pruebas piloto con grupos reducidos.** Estas pruebas piloto se emplean en fases de desarrollo más avanzadas, una vez que el formato del sistema y sus características han comenzado a estabilizarse. Para realizar la/s prueba/s se habilita uno o más laboratorios. El proceso consiste en escoger una muestra reducida pero representativa de usuarios reales del sistema, e interrogarlos antes y después de emplear el mismo. En aplicaciones educativas, mediante este tipo de experimento se puede ver si se han entendido o aprendido aspectos concretos de los contenidos o del uso de la aplicación, lagunas o puntos débiles en la instrucción/aprendizaje o si hay patrones de errores instruccionales (Mark y Greer, 1993). (Tessmer, 1993) añade que las pruebas piloto con grupos reducidos en ocasiones se emplean para validar los cambios inducidos como resultado de pruebas piloto uno a uno y por revisiones de expertos.
- **Pruebas de campo.** En dichos experimentos el diseñador o evaluador se traslada hasta el entorno de trabajo de los usuarios finales y observa el sistema en acción durante algún tiempo, es decir, sin habilitar un sitio nuevo específico para realizar el experimento, de manera que éste se efectúa en las condiciones de trabajo más reales posibles y con un número considerable de usuarios-participantes. Este tipo de pruebas involucra a una gran cantidad de personas, tanto participantes, como administradores y evaluadores, por lo que el coste asociado a la prueba es considerable. Consecuentemente, es vital que tanto los participantes como los lugares escogidos sean buenos representantes de la población total. (Tessmer, 1993) señala que el objetivo de estas pruebas es hacer revisiones en situaciones prácticamente reales, siendo especialmente útil para identificar problemas y resultados imprevistos que pueden surgir como consecuencia de la introducción de una aplicación nueva en un entorno de trabajo concreto, donde coexisten otra serie de aplicaciones con sus condiciones específicas. Así, según (Schofield, Evans-Rhodes et al., 1990), la evaluación puede ayudar a los investigadores en la exploración y prevención de resultados inesperados. Para sacar el mayor provecho a las pruebas, suele ser interesante que la propia aplicación (cuando es una aplicación informática) incluya software específico para capturar y registrar los

errores, los comandos utilizados, la frecuencia de consulta de las ayudas en línea así como para calcular medidas de productividad.

- **Test beta.** Son otro tipo de prueba piloto, cuya referencia, en esta ocasión, ha sido recogida de (Nielsen, 1993) y (Shute y Regian, 1993). En estas pruebas, la compañía desarrolladora del software selecciona un grupo pequeño de clientes para que estos aporten comentarios de los productos que en breve lanzarán al mercado. Los comentarios son el feedback necesario para retocar y elaborar el producto definitivo. El cauce de recogida de la información desde los usuarios reales puede ser, entre otros, vía línea directa telefónica, correo electrónico, dirección URL u opción específica en el menú de la aplicación. En (Welch y Brownell, 2000) se recoge la evaluación empírica de un producto comercial que estaban a punto de comercializar en esas fechas, y para el que empleaban, entre otras técnicas para evaluar el producto, el test beta.

Capítulo 4
El paradigma experimental en la Ingeniería del Software

La Ingeniería del Software ha adoptado el paradigma experimental como instrumento para la investigación. No basta sólo con medir la calidad de los productos software, sino que también es necesario aplicar métodos empíricos en la evaluación y validación de resultados. En este capítulo se argumenta la necesidad de la experimentación como mecanismo para validar nuevos procesos, métodos o herramientas, para seguidamente presentar los principios fundamentales de la experimentación en el área de la ingeniería del software. De cara a comparar resultados, se enuncian procedimientos para la normalización de datos experimentales y los tests de contraste más utilizados.

4.1. La necesidad de la experimentación empírica

La importancia de la experimentación en la ingeniería del software viene dada por la necesidad de probar la mejoría que introduce el uso de una nueva técnica o herramienta. Igual que ocurre en la ciencia en general, y en la ingeniería en particular, una teoría no se acepta por el prestigio del proponente ni por el mero hecho de ponerla de moda, sino que debe ir avalada por un conjunto de pruebas experimentales que confirmen su fiabilidad, su utilidad y/o su eficiencia.

La experimentación pretende suavizar el problema de decantarse por una alternativa relegando las otras. Esto no quiere decir que la ingeniería

del software sea igual que otras ciencias experimentales como la biología o la química. La principal diferencia estriba en la intervención del factor humano. En la ingeniería del software los sujetos experimentales son siempre personas, con su habilidad, inteligencia, experiencia, motivación (o falta de la misma) y, también, con sus días buenos y malos. Esto hace que las condiciones experimentales no sean siempre uniformes y que puedan perturbar los resultados obtenidos. Para disminuir este riesgo han surgido un conjunto de teorías y técnicas prácticas heredadas de la experimentación en otras disciplinas que se agrupan bajo el nombre de diseño experimental.

Existe constancia de algunos estudios empíricos aislados en la ingeniería del software (Basili and Turner 1975; Belady and Lehman 1976; Shneiderman, Mayer et al. 1977). Basili fue quien puso de manifiesto la necesidad de experimentar y mejorar en dicha disciplina, al considerar la ingeniería del software como un laboratorio de ciencia (Basili, 1985; Basili, Selby et al., 1986). Posteriormente, se han publicado otros artículos que han seguido apoyando esta necesidad de contrastar empíricamente, por ejemplo, métodos de análisis y diseño de desarrollo de software (Basili 1996; Kitchenham 1996; Zelkowitz and Wallace 1998; Pfleeger 1999), el grado de adaptabilidad de la aplicación al usuario (Brusilovsky y Pesin, 1998) o la complejidad de la navegación a través de los enlaces (Höök y Svensson, 1999).

Paulatinamente, el paradigma experimental en la ingeniería del software se está consolidando y las propuestas son más formales. Prueba de ello son las publicaciones recientes de varios libros de introducción al proceso experimental como, por ejemplo, (Dolado y Fernández, 2000; Juristo y Moreno, 2001; Wohlin, Runeson et al., 2000). Tal y como explican Juristo y Moreno, existen dos tipos de trabajos de investigación de la ingeniería del software experimental. El primer tipo consiste en proponer alguna herramienta o método y probar mediante experimentación que su aplicación mejora algún aspecto del desarrollo del software. A veces, y aunque pudiera parecer lo contrario, este trabajo resulta demasiado amplio para ser abordado en una única tesis doctoral. Debido a esto, ha surgido el segundo tipo de trabajos que se dedican a desarrollar un conjunto de experimentos con rigurosidad que prueba la eficiencia de algún método, técnica, lenguaje, proceso, etc. frente a otra propuesta. Ejemplos de investigaciones incluyendo familias de

experimentos son (Rolón, García et al., 2007; Tuya, Ramos et al., 2007) o bien las tesis doctorales (Genero 2002; Otero 2003).

El objetivo de la experimentación es proporcionar datos experimentales, en oposición a los datos obtenidos por observación de la actuación de las unidades elementales (individuos) de una muestra o población. Sin embargo, no es viable registrar todos los datos que pudieran generarse en torno a un experimento, por lo que es necesario el discernir aquello que realmente tiene valía para el experimento en cuestión.

El siguiente apartado recoge someramente una introducción de las variables que intervienen en los procesos experimentales junto con sus características principales y las repercusiones de la elección de una frente a otras en el análisis e interpretación de los resultados del experimento.

4.2. Las variables del proceso experimental

La variable o factor de un proceso experimental puede definirse como cualquier atributo de los objetos o seres que sea medible y cuyos valores varían (por ejemplo, color, magnitud, peso, etc.). Existen dos alternativas ampliamente empleadas para determinar los tipos de las variables:

- Dependiendo de la naturaleza de las **variables** se distinguen dos tipos de variables: las cualitativas y las cuantitativas. Las **cualitativas** son aquellas cuyas categorías o niveles no se pueden ordenar con respecto a la magnitud –religión, profesión, etc. En contraposición, las **cuantitativas** se pueden ordenar, pudiendo ser **continuas** o **discretas**.
- De acuerdo con los intereses u objetivos del investigador, existen variables dependientes e independientes. La **variable dependiente** (o variable *explicada*) es cualquier aspecto de la conducta medido por el experimentador para evaluar los efectos de la variable independiente manipulada, y que interesa estudiar al investigador (por ejemplo la velocidad, la frecuencia, el coste, etc.); mientras que la **variable independiente** es cualquier variable manipulada por el investigador, bien directamente o por medio de selección para determinar su efecto

en la variable dependiente que, a veces se conoce también como variable de *estado* o *explicativa*.

La **escala de medición** de las variables experimentales es una relación entre el sistema empírico y el numérico. Este concepto es distinto al **tipo de escala**, que engloba a todas las escalas que permiten la misma transformación admisible. Según (Zuse, 1998) una **transformación admisible** es una operación matemática que permite la conversión de la escala garantizando la conservación de la condición de representación. Las transformaciones más ampliamente empleadas son el cambio de origen y el cambio de escala o de unidad (véanse, respectivamente, los ejemplos en la Ecuación 1). No obstante, la cuestión de la transformación de la escala no es tan simple, ya que puede depender del contexto de aplicación, no resultando ser siempre válida. En (Wohlin, Runeson et al., 2000) se aborda el tema con ejemplos ilustrativos y el apartado 4.5 prosigue con el aspecto de la transformación de escalas pero desde el punto de vista de unificación de métricas para hacerlas comparables.

Asimismo, las tareas de identificar y formalizar las escalas tampoco son, en absoluto, triviales. Más bien todo lo contrario, dado que suelen desarrollarse con muchas incertidumbres. Para facilitarlas en (Zuse, 1998) se han propuesto algunos marcos formales. Aunque los investigadores destacados en el área insisten en la necesidad de abordar dichas tareas para dar credibilidad a las propuestas, también conceden cierto margen de confianza al amparo de que la ciencia no es matemática (Tukey, 1986).

$$Y = a + X \qquad \text{cambio de origen}$$

$$Y = b\,X \qquad \text{cambio de escala o de unidad}$$

Ecuación 1.- Transformaciones lineales de escalas

La importancia de identificar claramente las variables de un experimento junto con las escalas de la medida tiene gran importancia, ya que condiciona el diseño del experimento y viceversa. Dependiendo de la escala, tendrán sentido o no determinadas operaciones sobre los valores medidos (de acuerdo con las transformaciones admisibles): el tipo de escala determina qué estadísticos son apropiados para el análisis de los datos y presentación de los resultados. En (Wohlin, Runeson et al., 2000) se muestra la Tabla 1, la cual resume la relación escala-estadístico apropiado:

Tipo de escala	Medidas de tendencia central	Medidas de dispersión	Medidas de dependencia
Nominal	Moda	Frecuencia	
Ordinal	Mediana, percentil	Intervalo de variación	Coef. correl. Spearman Coef. correl. Kendall
Intervalo	Media	Desviación típica, varianza y rango	Coef. correl. Pearson
Ratio	Media geométrica	Coef. de variación	

Tabla 1.- Tipo de escala vs. Estadísticos apropiados

4.3. Principios básicos del diseño experimental

El diseño estadístico de experimentos es el proceso de planificar un experimento para obtener datos apropiados, que pueden ser analizados mediante métodos estadísticos, con objeto de producir conclusiones válidas y objetivas. Se requiere de un enfoque estadístico del diseño de experimentos para obtener conclusiones significativas a partir de los datos. La metodología estadística es el único enfoque objetivo para analizar un problema que involucra datos sujetos a errores experimentales. Es por ello que hay dos aspectos a ser considerados en cualquier problema experimental: el diseño del experimento y el análisis estadístico de los datos. Ambos están estrechamente relacionados, ya que el método del análisis depende directamente del diseño empleado tal y como se ha anticipado en la sección anterior.

En la relación causa-efecto que la variable independiente ejerce sobre la dependiente pueden mediar otras fuentes extrañas de variación. Algunas de estas fuentes de error se deben a la variabilidad que existe entre los participantes, la diferencia de concentración debido a la prolongación del experimento, los fallos técnicos o problemas de sincronización en el funcionamiento de los ordenadores y de la red, etc. (Nielsen, 1993). Como, en realidad, la eliminación completa no es posible, hay que intentar disminuir esta variabilidad controlando tantas variables como sea posible. Mediante la aplicación de **técnicas de control**, tales como la *asignación aleatoria*, el *bloqueo* y el *equilibrado*, se logra **minimizar el error experimental**. Otros principios de diseño, como son la *replicación* (o

reproducción) y la *selección aleatoria*, permiten **asegurar la validez de los resultados**.

La réplica. Se refiere a la repetición del experimento básico. Para reproducir el experimento de la forma más estricta posible es preciso generar kits de laboratorio con todo el material experimental. No obstante, la réplica de un estudio empírico no implica necesariamente la repetición bajo las mismas condiciones. En (Basili, 1999) se pueden encontrar los diferentes tipos de réplicas que se pueden ejecutar en ingeniería del software clasificadas por: réplicas estrictas, réplicas que varían las variables independientes o dependientes y réplicas que cambian el contexto de experimentación. Sirva como ejemplo el experimento concebido en (Porter, Votta et al., 1995), ya que tras el experimento original y gracias el kit experimental proporcionado por los investigadores, con posterioridad se han podido realizar las réplicas independientes (Fusaro, Lanubile et al., 1997) y (Miller, Wood et al., 1998).

La realización de réplicas posee dos propiedades importantes. En primer lugar, permite al experimentador obtener una estimación del error experimental. Esta estimación del error se convierte en una unidad de medición básica para determinar si las diferencias observadas (obtenidas empíricamente) en los datos son, en realidad, estadísticamente significativas. En segundo lugar, la realización de réplicas permite al experimentador calcular una estimación más precisa del efecto de un factor en el experimento. Por ejemplo, si se usa la media de la muestra ($\bar{x}$) como estimación de dicho efecto, y hay *n* réplicas, la varianza de la media muestral es $s_{\bar{x}}^2 = \frac{\sum (x_i - \bar{X})^2}{n-1}$ (Wonnacott y Wonnacott, 1991).

La consecuencia práctica de esto es que, si no se hacen réplicas, las inferencias realizadas para los datos no serán satisfactorias, pues las diferencias observadas podrían bien deberse al error aleatorio inherente de la propia medición experimental y, por lo tanto, no ser diferencias sustanciales, significativas. Por otra parte, si *n* es suficientemente grande y el error experimental es razonablemente pequeño, las diferencias observadas quedan claramente definidas. Por todo ello, uno de los aspectos más importantes a la hora de diseñar un experimento es estimar el *tamaño de la muestra necesario* para alcanzar una potencia aceptable y controlar la probabilidad de cometer un error de tipo I (sección 4.4.2.3).

La aleatorización. Es la piedra angular en la se que fundamenta el uso de los métodos estadísticos en el diseño experimental. La aleatorización se puede producir a dos niveles distintos, a saber, la selección aleatoria y la asignación aleatoria. En el **muestreo aleatorio** (en inglés, *random sampling*) cada uno de los grupos debe ser una muestra (representativa) extraída al azar de la población. Se utiliza especialmente en encuestas. En cuanto a la **asignación aleatoria** (en inglés, *random assignment*) la aleatorización se refiere a que los grupos de sujetos correspondientes a los distintos niveles de la variable independiente se forman mediante asignación aleatoria. Ésta es la técnica de aleatorización que se emplea principalmente en los métodos experimentales, puesto que extraer los sujetos aleatoriamente de "toda" la población es inviable. Aunque interesa la representatividad del colectivo, la principal preocupación de los experimentadores es mostrar la influencia de la variable independiente en la dependiente, y, para ello, asumen que esta influencia es independiente de variables tales como la localización geográfica de los individuos o bien el centro donde trabajan o estudian. Esta técnica constituye una forma de controlar las variables extrañas, por equilibrado, que están presente en los sujetos: inteligencia, experiencia, etc.

El análisis por bloques. Es una técnica de control que se usa para incrementar la precisión del experimento. Un bloque es una proporción del material experimental que es más homogénea que el total del material. Técnicamente consiste en organizar grupos emparejados (*matched*), denominados bloques, atendiendo a una variable relacionada con la variable dependiente, para posteriormente realizar un análisis por bloques donde se harán comparaciones entre las condiciones de interés del experimento dentro de cada uno de los bloques.

El equilibrado. Se dice que el diseño experimental está equilibrado cuando cada grupo tiene el mismo número de individuos. Aún habiendo diseñado un experimento con grupos equilibrados, pueden suceder diversos contratiempos que causen desequilibrio. Por ejemplo, el abandono de los individuos en las sucesivas sesiones de ejecuciones del experimento de (Laitenberger y DeBaud, 1997) ocasionó que en cada celda no hubiese información sobre el mismo número de personas, lo que complicó el análisis estadístico de los resultados.

4.4. Directrices generales del proceso experimental

Un proceso experimental puede entenderse como una evaluación empírica, aunque no todo proceso experimental puede ser considerado como experimento científico controlado. De hecho, y al igual que construir software requiere un proceso de desarrollo compuesto por un conjunto de actividades (tales como, análisis, diseño e implementación, entre otros), conducir un experimento implica seguir varios pasos en un orden establecido. Lógicamente, estos pasos engloban el proceso experimental necesario para que un experimento controlado proporcione resultados útiles y significativos (Fenton y Pfleeger, 1998; Wohlin, Runeson et al., 2000). En la Figura 7 se muestran las fases que constituyen el proceso experimental propuesto por (Wohlin, Runeson et al., 2000). La fase de planificación se ha completado con la propuesta de (Tessmer, 1993).

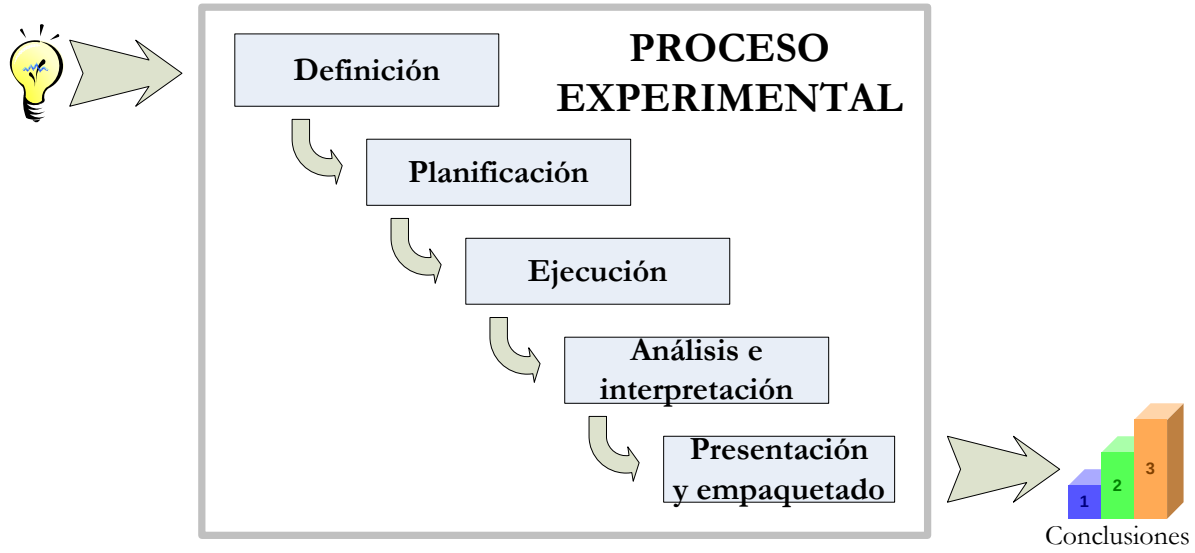


Figura 7.- Proceso experimental de Wohlin et al.

La concepción de una **idea** es el punto de partida de todo proceso experimental. A partir de dicha idea, la primera fase es la **definición** clara y concisa del objetivo del estudio empírico. Mientras que la fase de definición del experimento determina el porqué de la realización del experimento, la fase de planificación debe organizar cómo debe ejecutarse éste. En esta ocasión, la fase de **planificación** comprende el estudio logístico de diversas tareas implicadas: la selección de las variables, el planteamiento de la hipótesis (si procede), la selección de los sujetos involucrados, la elección del diseño en función de las variables e

hipótesis, la preparación de los materiales y la realización de pruebas piloto (si procede). Para que la **ejecución** del estudio sea correcta, ésta debe ajustare al plan y diseño experimental. Los datos recogidos serán filtrados y examinados en la fase de **análisis** mediante el contraste estadístico adecuado. Seguidamente, en la fase de **interpretación** se explicarán los resultados, teniendo en cuenta el rechazo o aceptación de las hipótesis planteadas en la etapa de planificación. Para finalizar, la fase de **presentación** incluirá la documentación tanto de los resultados (por ejemplo, mediante informes técnicos) para los diversos sujetos interesados como de los materiales experimentales, con el fin de crear kits de laboratorios fácilmente replicables. Evidentemente, tras la concepción de la idea y definición y ejecución del proceso experimental, el objetivo final es la obtención de **conclusiones** fiables.

En los siguientes apartados se profundiza en las fases más importantes identificando y concretando el desglose de las actividades internas.

4.4.1. Definición del objetivo

Aunque parezca una obviedad, la identificación y enunciación clara del problema no es sencillo. El enfoque GQM (Goal/Question/Metric) ofrece una alternativa sencilla para formular una correcta definición del objetivo (van Solinger y Berghoout, 1999) mediante la especificación de 5 parámetros incluidos en una plantilla (Figura 8) y que son: el objeto de estudio, el objetivo o propósito del estudio, el aspecto de calidad que se estudiará, desde qué perspectiva se va a realizar el estudio y, finalmente, en qué contexto se va a realizar el experimento.

Analizar <*objeto(s) de estudio*>
con el fin de <*propósito*>
con respecto a su <*aspecto de calidad*>
desde el punto de vista del <*perspectiva*>
en el contexto de <*entorno*>

Figura 8.- Plantilla GQM para la definición del objetivo

Para determinar los valores de los mencionados 5 parámetros se pueden emplear las propuestas de la Tabla 2 recogidas desde (Wohlin, Runeson et al., 2000):

Objeto(s) de estudio	Propósito	Aspectos de calidad	Perspectiva	Entorno
Lenguaje	Caracterizar	Coste	Cliente	Sujetos:
Métrica	Controlar	Efectividad	Desarrollador	Estudiantes
Modelo	Evaluar	Mantenibilidad	Gestor del proyecto	Profesionales
Proceso	Mejorar	Comprensibilidad	Investigador	Objetos
Producto	Monitorizar		Programador	Complejidad
Teoría	Predecir		Usuario	Dominio
	Validar			

Tabla 2.- Ejemplos de los parámetros de la plantilla GQM

Con independencia de emplear o no el método GQM, el conocerlo sirve para dar una idea de lo que se entiende por detallar el objetivo del experimento.

4.4.2. Planificación

Definido el objetivo del experimento, es el momento de realizar la planificación del mismo, lo que abarca concretar los siguientes aspectos: la *selección de las variables y de los sujetos participantes*, la *formulación de la hipótesis* y del *diseño* (experimental) del proceso, la *instrumentación* a emplear, y la realización de *pruebas piloto*. Dichos aspectos se amplían a continuación.

4.4.2.1. Selección de las variables

Ello incluye la selección de las variables dependientes e independientes que intervienen y relacionan el objetivo con la hipótesis del experimento ya definido. En el supuesto de haber empleado la plantilla GQM, los factores que caracterizarían el <*objeto de estudio*> constituyen las variables independientes, mientras que las variables dependientes miden el <*aspecto de la calidad*>. Junto con la selección de las variables hay que concretar la escala de medida apropiada (Fenton y Pfleeger, 1998).

Es conveniente ser especialmente cuidadosos a la hora de delimitar las variables que se van a considerar durante el desarrollo del experimento ya que, tal y como se ha indicado anteriormente, existe otro tipo de variables cuya forma de control influye en la elección del diseño experimental y en el resultado del experimento. Así, por ejemplo, la determinación de recopilar información a través de la administración de

cuestionarios electrónicos frente a los tradicionales en formato de papel incide tanto en el tiempo de completado de la encuesta por el individuo encuestado así como en la respuesta aportada por el mismo.

Además, es preciso discernir y descartar factores superfluos que no tienen interés en el experimento, debido a que, de por sí, en torno a los experimentos se genera una cantidad ingente de datos que deben ser tratados.

4.4.2.2. Selección de los sujetos participantes

(Tessmer, 1993) distingue dos tipos de sujetos involucrados, a saber, los activos y los pasivos. En este punto se trata de establecer los sujetos activos que se encargan de llevar el proceso adelante. Por tanto, entre sus competencias estarán el diseñar, planificar, conducir, ejecutar, analizar, gestionar y supervisar todas las actividades que se precisen realizar durante el ciclo de vida del experimento. Igualmente, deben desvelarse quiénes serán los sujetos pasivos, esto es, aquellos con los que contactarán los sujetos activos para recabar información. Una vez conocido el perfil de ambos tipos de participantes, se debe concretar el modo de captar a unos y otros.

En el área la ingeniería del software conseguir suficientes sujetos participantes para la experimentación puede resultar un impedimento difícil de superar. En aquellas disciplinas donde se estudia el comportamiento humano, como puede ser la sociología, es fácil disponer de sujetos puesto que, normalmente, no se requiere ninguna experiencia previa. Por el contrario, la ingeniería del software está limitada en cuanto a recursos humanos se refiere. Ello se debe a que sólo un pequeño porcentaje de la población posee el conocimiento o la capacidad necesaria para desarrollar, evaluar o mejorar software, y en consecuencia ser apto para la experimentación en este campo. El obstáculo es aún mayor cuando el experimento o prueba precisa de individuos especializados en un determinado campo. Y aunque la participación remunerada puede incentivar a la participación a los desmotivados, cuando los requisitos que deben de cumplir los participantes son muy restrictivos, no hay forma posible de suplir la carencia de individuos para tomar parte en el estudio.

Esas y otras razones influyen en que el número de individuos pasivos experimentales que participan en los estudios empíricos no sea siempre

elevado. Justamente en los estudios empíricos de (Basili, Green et al. 1996; Laitenberger and DeBaud 1997; Cartwright 1998; Zendler, Pfeiffer et al. 2001) y (Otero, 2003) participaron entre 10 y 30 sujetos en cada uno. Estas cifras contrastan con las de los estudios de (Finney, Rennolls et al., 1998) y (Caro, 1988) en los cuales participaron 147 y 1911 participantes respectivamente.

4.4.2.3. Formulación de la hipótesis. Estimación puntual y por intervalos

En numerosas ocasiones, las preguntas que el experimentador quiere responder se pueden modelar mediante la denominada prueba de hipótesis. En dichos casos, el planteamiento de la hipótesis consiste en transformar la idea concebida para el experimento en una sentencia formal. En un contraste de hipótesis lo que conviene al experimentador es rechazar la llamada *hipótesis nula (*H_0*)* (Fusaro, Lanubile et al., 1997). Con frecuencia los investigadores enuncian como sus hipótesis lo contrario de lo que consideran verdadero, con la esperanza de que los procedimientos de demostración (y los datos) conduzcan a rechazarlos, siendo éste el modo habitual en la que el experimentador demuestra que su propuesta es mejor que otras. Por el contrario, en determinados estudios lo que el investigador pretende verificar es la igualdad de las propuestas (de tratamientos alternativos, de procesos, etc.). A estos estudios se les suele denominar **estudios de equivalencia** (también **negativos** o bien de prueba **de hipótesis de nulidad)**.

De acuerdo con lo anterior, al definir "la hipótesis" realmente lo que se tiene que formular son dos hipótesis, las denominadas *Hipótesis Nula* e *Hipótesis Alternativa*:

- **Hipótesis nula** (denotada como $\mathbf{H_0}$). Generalmente se plantea como hipótesis nula la hipótesis de que no existe diferencia, es decir, que la diferencia es nula entre los valores a comparar. Dicho de otra manera, en (Dolado y Fernández, 2000) se explica que la hipótesis nula establece que no hay diferencias entre dos o más tratamientos (esto es, entre dos o más métodos, procesos, herramientas u otras condiciones cuyos efectos se quieran medir) con respecto a la/s variable/s dependiente/s que se mide/n. Por lo general, la hipótesis nula se plantea de tal modo que especifique un valor exacto del parámetro.

- **Hipótesis alternativa** (denotada como $\mathbf{H_1}$, o H_A). Es aquella hipótesis contra la cual se quiere contrastar la hipótesis nula, luego, es aquella que sostiene que las diferencias no son nulas. Consecuentemente, el rechazo de la hipótesis nula siempre conduce a la aceptación de la hipótesis alternativa, y viceversa.

Para determinar si hay o no diferencias, se realiza un contraste o **prueba de hipótesis**, que es un procedimiento estadístico mediante el cual se investiga la aceptación o el rechazo de la hipótesis nula con base a los resultados muestrales. Los procedimientos de prueba de hipótesis dependen del empleo de la información contenida en una muestra aleatoria de la población de interés. Si esta información es consistente con la hipótesis, se concluye que ésta es verdadera; sin embargo, si esta información es inconsistente con la hipótesis, se concluye que ésta es falsa.

El procedimiento estadístico más extendido para realizar una prueba de hipótesis es emplear una prueba o **test de contraste** específico como, por ejemplo, la t de Student, el test de Kruskal-Wallis, la correlación de Spearman, o ANOVA. En la sección 4.6 se recogen algunos de los tests de contraste más empleados. Conviene apuntar que cada test tiene sus propias condiciones de aplicabilidad y que la estimación que realiza es puntual. Para rechazar o aceptar la prueba de hipótesis el valor crítico usual empleado es 0,05 y se denomina *valor de p* o *nivel de significación estadístico*. Así, cuando el valor de p es inferior al umbral ($p<0,05$), se interpreta como que hay suficiente evidencia estadística para descartar la hipótesis nula.

Otro procedimiento alternativo para realizar la prueba de contraste es el uso de intervalos de confianza. El **intervalo de confianza** (IC) da el margen de valores en los que *es previsible esperar que se encuentre la verdadera diferencia* entre los dos tratamientos o procesos comparados, para una probabilidad dada (habitualmente el 95%) (Wackerly, 2002). Consecuentemente, si el estadístico de prueba cae dentro de un intervalo de valores "comunes" descritos por la distribución nula, el estadístico de prueba se considera consistente con la H_0, que entonces no se rechaza.

Hay que comprender que la aceptación de una hipótesis simplemente implica que los datos obtenidos no dan suficiente evidencia para rechazarla. Por otro lado, el rechazo de una hipótesis implica que la evidencia muestral refuta la hipótesis planteada. Puesto de otra manera,

el rechazo de una hipótesis significa que existe una pequeña probabilidad de obtener la información muestral observada, cuando realmente dicha hipótesis es verdadera. En consecuencia, resuelto el contraste, se pueden dar las siguientes situaciones:

- Se acepta H_0, cuando ésta es realmente cierta.
- Se rechaza H_0, cuando ésta es realmente cierta; luego, se rechaza incorrectamente cometiendo un error, el designado por **error de tipo I** de un contraste de hipótesis. La probabilidad de que el error de tipo I ocurra viene dado por el valor de α y se conoce como el **nivel de significación** del test.
- Se acepta H_0, cuando ésta es falsa; luego se acepta incorrectamente, cometiendo un error, el designado por **error de tipo II** y la probabilidad de que ello ocurra se representa con β.
- se rechaza H_0, cuando ésta es realmente falsa, cuya probabilidad es (1-β) y se le llama **potencia del contraste estadístico**.

Convencionalmente, se asume que con un α de 0,05 y un β de 0,02 se logra un equilibrio conveniente entre los errores de tipo I y II (Cohen, 1992). En términos estadísticos, la potencia es (1- β), y el nivel ideal de potencia debiera ser siempre igual o superior a 1-0,2. Esto es, en una escala que va de 0 a 1 el nivel mínimo de potencia requerido en una investigación cuantitativa según Cohen es de 0,8.

Lógicamente, al tomar una decisión estadística sobre una prueba se está asumiendo un cierto riesgo, ya que se pueden cometer dos tipos de errores: rechazar una hipótesis que es válida o aceptar una hipótesis que es falsa. Esto, junto con las limitaciones e interpretaciones erróneas del uso de *p* han motivado muchas críticas sustentadas empíricamente (Borges, San Luis et al. 2001; Chalco 2003; Newcombe and Merino-Soto 2006). Por ejemplo, (Schmidt, 1996) aboga por el abandono de los estudios de significancia estadística, mientras que otros autores más conciliadores no descartan su uso, pero aconsejan que las estimaciones puntuales vayan acompañadas de cantidades adicionales (Mulaik, Raju et al., 1997). En consecuencia, progresivamente, revistas científicas han dejado de admitir artículos originales en los que sólo aparece ese parámetro sin incluir otros que hoy se consideran más apropiados y menos sujetos a interpretaciones erróneas, tales como el intervalo de confianza o la magnitud del efecto (*effect size*). Así lo recogen

específicamente en sus respectivas directrices de publicación, por ejemplo, la *British Medical Journal* (desde hace más de década), la *Language learning* (en instrucciones para autores del 2002) y la *Asociación Psicológica Americana* (en su manual de publicación de 2001).

4.4.2.4. Elección del diseño experimental

Una vez aclarado el contexto de partida, establecidos los objetivos del experimento y concretadas las variables junto con la hipótesis, es el momento de concretar el diseño respetando los principios básicos del diseño experimental enunciados en la sección 4.3. La elección del diseño implica (1) determinar el tamaño de la muestra (o número de réplicas a realizar) y grupos muestrales, (2) concretar la selección aleatoria de los elementos muestrales y (3) organizar la logística de las conducciones de los ensayos experimentales.

4.4.2.5. Instrumentación a emplear

Los instrumentos de un experimento son de tres tipos, a saber, objetos, guías e instrumentos de medida. Los **objetos** del experimento pueden ser, entre otros, documentos de la especificación o del código. Las **guías** son necesarias para dirigir a los participantes en el experimento y pueden incluir, por ejemplo, descripciones del proceso y listas de comprobación. Si el experimento fuera para comparar métodos alternativos, habría que elaborar las guías de cada uno de los métodos. En ocasiones, además, es preciso también entrenar a los participantes en los métodos que utilizarán. Las **medidas** de un experimento se obtienen a partir de los datos recopilados durante la conducción del mismo. En los experimentos intensivos humanos, los datos se recogen generalmente de forma computerizada, manualmente o bien mediante entrevistas. En consecuencia, durante la planificación del estudio habrá que diseñar y preparar el material de los cuestionarios y preguntas, y validarlos con gente que tenga habilidades y formación similar a la población objetivo del ensayo.

El objetivo general de la instrumentación es proporcionar los medios para realizar el experimento y supervisarlo, sin que ello repercuta al control del experimento. Los resultados deberán ser los mismos con independencia de cómo se equipa el estudio. Si la instrumentación afecta al resultado, los resultados serán inválidos. En la sección 3.3 se han

presentado técnicas para recopilar datos y la siguiente sección ahonda en la validez de un experimento.

4.4.2.6. Validez experimental

Pudiendo estudiarse la validez del experimento con mayor o menor rigor, en este trabajo, y de acuerdo con (Wohlin, Runeson et al., 2000), se entiende por validez adecuada aquella cuyos resultados son válidos para el interés de la población de la cual se han derivado los resultados e, igualmente, son válidos para la población para la que se quieren generalizar. Y es que, relacionada con los resultados del experimento, una cuestión obvia y fundamental es hasta qué punto son estos válidos.

La cuestión de la validez no debe limitarse a la fase de los resultados sino que debe remontarse a la planificación del mismo. Mediante un buen diseño experimental, se deberán maximizar las propiedades de los experimentos formales. Además, para ello es necesario minimizar o amortiguar la presencia de los factores que amenazan la validez de los resultados experimentales.

A continuación, y de acuerdo con la categorización propuesta por (Cook y Campbell, 1979), se enuncian las propiedades de los experimentos formales junto con aquellas amenazas más relevantes en este ámbito de experimentación empírica:

- **Validez de la conclusión o fiabilidad**. Se dice que un experimento es fiable cuando, al repetirse, se obtienen los mismos resultados. Esta validez se refiere a la relación entre el tratamiento y los resultados. Por ello es preciso el control y la replicación, bien interna o externa, del experimento. Así, algunas amenazas relevantes que pueden malograr el experimento son la baja potencia estadística, la violación de supuestos de los contrastes estadísticos o la heterogeneidad de los sujetos.
- **Validez constructiva**. Es el grado con que tanto las variables independientes como las dependientes reflejan o miden de manera certera las hipótesis del experimento. Admite también la formulación del grado en que se ajusta el marco del experimento a la realidad bajo estudio. En este caso, las amenazas a atenuar pueden tener origen social o de diseño. Por ejemplo, el número de variables consideradas en el experimento puede ser insuficiente para reflejar la realidad, el

grado de precisión con el que se ha descrito el objetivo del experimento puede ser insuficiente y la aprensión a ser evaluados o las propias expectativas del experimentador pueden condicionar los resultados.

- **Validez interna**. Es el grado con el que se puede atribuir que los cambios de la variable dependiente son debidos a la influencia de la variable independiente que se estudia. Algunos factores que pueden influir en la validez de la prueba son: la habilidad de los individuos de la muestra, la madurez o aprendizaje de los mismos durante el desarrollo del experimento, su tasa de abandono, el propio diseño del experimento y eventos específicos que pudieran alterar los resultados.
- **Validez externa**. Es el poder de generalización de los resultados obtenidos en el laboratorio a las condiciones normales. Normalmente, es consecuencia de haber seleccionado una muestra no representativa de la población a la que se quieren generalizar los resultados. Esto desencadena en una interacción entre la selección de los individuos y el tratamiento, y puede llegar a invalidar la extrapolación de los resultados.

Las personas interesadas en conocer más factores que hacen peligrar la validez de los experimentos controlados pueden consultar, entre otros, los libros (Kitchenham, Pickard et al., 1995) y (Wohlin, Runeson et al., 2000). En un principio, (Campbell y Stanley, 1973) definen una primera lista de "12 amenazas" a la validez interna y externa de un experimento. Más adelante, (Cook y Campbell, 1979) amplían dicha lista, añadiendo amenazas a los cuatro tipos de validez experimental citados anteriormente.

4.4.2.7. Pruebas piloto

Previamente a la ejecución del experimento, los investigadores se deben de plantear si es preciso que los sujetos participantes asistan primero a una etapa de formación o de entrenamiento que les permita familiarizarse con los materiales de experimentación. En colación con este asunto, (Tessmer, 1993) apunta la necesidad de realizar varias pruebas piloto a fin de comprobar la idoneidad del diseño de los materiales y de las pruebas y localizar posibles deficiencias o carencias en los mismos. La sección 3.3.4 ha presentado los tipos y características principales de las pruebas piloto, si bien para mayor detalle puede

consultarse (Tessmer, 1993). Dicho autor añade, además, que éstas deben repetirse en tanto en cuanto se sigan detectando problemas. Específicamente, las pruebas piloto proporcionan información acerca de la consistencia del material experimental, permiten realizar una comprobación del sistema de medición, pueden dar una idea aproximada del error experimental y dan la oportunidad de poner en práctica la técnica experimental global. Así mismo, ofrece la oportunidad para revisar, de ser necesario, las decisiones tomadas hasta el momento.

4.4.3. Ejecución

La ejecución de un experimento controlado consiste en la conducción de los ensayos o pruebas ciñéndose al diseño planificado. Los errores en el procedimiento experimental en esta etapa destruirán por lo general la validez experimental. La realización individual de cada uno de los ensayos de un experimento puede limitarse a unos minutos o bien sobrepasar el día. Del mismo modo, la ejecución de todos los ensayos que conforman el experimento puede extenderse durante un periodo de tiempo variable, desde unos días hasta incluso abarcar varios años. Consecuentemente, es de importancia capital registrar la evolución de la ejecución así como llevar un registro detallando las incidencias acaecidas y la resolución de las mismas.

4.4.4. Análisis e interpretación de los datos

Para analizar e interpretar los datos recopilados durante la ejecución de las pruebas hay que comenzar intentando entender los mismos, lo que se puede hacer mediante técnicas de la estadística descriptiva, a fin de que los resultados y las conclusiones sean objetivos y no de carácter apreciativo. La estadística descriptiva puede emplearse para describir y representar gráficamente aspectos remarcables del conjunto de datos. Por ejemplo, en la escala de medición, por dónde se hallan los datos, por dónde se concentran y por dónde se dispersan. El objetivo de la estadística descriptiva es proporcionar una idea de la distribución de los valores del conjunto, para lo cual se emplean medidas de tendencia central, de dispersión y de dependencia. En la tabla Tabla 1 se han mostrado algunos de los procedimientos más comunes. Para mayor

detalle se puede recurrir prácticamente a cualquier referencia de estadística de nivel universitario.

Un aspecto importante que hay que sopesar en esta fase es considerar si es o no preciso la reducción del número de las variables (por ejemplo, estudiando si existen variables que expliquen otras) así como el filtrado de los datos para descartar valores erróneos y valores extremos (*outliers*). La eliminación de outliers o valores imperfectos está relacionada con la validación de los datos y trata con la identificación de valores falsos de puntos obtenidos en la ejecución del experimento, lo cual incluye determinar si realmente los participantes pasivos han participado seriamente en el experimento. Para la depuración de outliers o datos imperfectos pueden consultarse (Frechtling y Sharp, 1997; Wohlin, Runeson et al., 2000). Sin embargo, además de eliminar los datos inválidos, puede interesar analizar si el volumen de información redundante es excesivo. Para ello, se estudia la viabilidad de reducir el número de variables mediante técnicas como el análisis factorial (o de variables) o bien el análisis de componentes, técnicas que pueden consultarse en (Kachigan, 1986; Manly, 1994).

Posteriormente, y debido a que muchas de las preguntas que el experimentador quiere responder pueden insertarse en el marco de la prueba de hipótesis, los procedimientos para probar hipótesis y estimar intervalos de confianza son muy útiles en el análisis de datos de un experimento diseñado. En la sección 4.6 se enumeran algunos de los procedimientos estadísticos más frecuentemente empleados a la hora de aceptar o rechazar la similitud de los resultados obtenidos en pruebas distintas, ya que la interpretación es un aspecto importante de esta fase y es la base de la toma de decisiones así como del objetivo de la realización del experimento.

Si el experimento se ha diseñado correctamente y se ha llevado a cabo de acuerdo con el diseño, la puesta en marcha de los métodos estadísticos necesarios no debe ser complicada, más aún hoy en día con el software estadístico disponible. A este tenor, la utilidad de los gráficos es doble, ya que pueden servir no sólo como sustituto a las tablas, sino que también constituyen por sí mismos una poderosa herramienta para el análisis de los datos, siendo en ocasiones el medio más efectivo no sólo para describir y resumir la información, sino también para analizarla. El propósito de un gráfico es similar a otra herramienta estadística: ayudar a

la comprensión y comunicación de la evidencia aportada por los datos respecto a una hipótesis de estudio. El gráfico científico debe servir por tanto para representar la realidad, no para generar nuevas realidades inexistentes fuera de la propia imagen. De este modo, y según la calidad de un gráfico estadístico, el objetivo de éste consiste en comunicar ideas complejas con precisión, claridad y eficiencia, de manera que: induzca a pensar más en el contenido que en la apariencia, no distorsione la información proporcionada por los datos, presente mucha información en poco espacio y favorezca la comparación de diferentes grupos de datos o de relaciones entre los mismos. No obstante, conviene puntualizar que la utilización de un escalado adecuado es fundamental sobre todo si se van a comparar diferentes gráficas. Entre los gráficos más empleados se hallan los gráficos de dispersión (o *scattered plots*, en inglés), los bigotes (también conocidos como cajas o *boxplots*, en inglés) e histogramas, pudiendo ver sus respectivos aspectos en la siguiente figura:

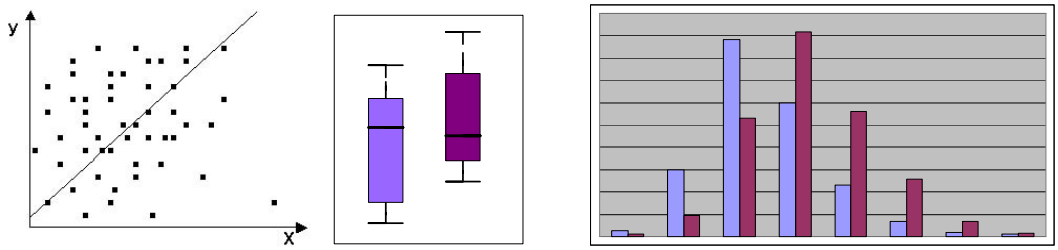


Figura 9.- Gráficos de dispersión, de bigote e histogramas

4.4.5. Presentación y empaquetado de resultados

Una vez analizados los datos, el experimentador debe extraer conclusiones prácticas acerca de los resultados, recomendar un curso de acciones a tomar y divulgar los mismos. Esto consiste principalmente en documentar los resultados bien para realizar publicaciones de investigación, bien como un paquete de laboratorio con vistas a posibilitar la realización de réplicas en un futuro o bien como informe donde se realiza el balance de la experiencia de alguna compañía. Es necesario elaborar una documentación comprensible y concienzuda de todo el proceso para alcanzar los objetivos fijados, al mismo tiempo que aportar sustento científico a las lecciones aprendidas.

No obstante, y para finalizar, hay que matizar que ni los experimentos zanjan la cuestión y que ni los métodos estadísticos pueden demostrar

que una(s) variable(s) posee(n) un efecto particular. Sólo proporcionan pautas generales en cuanto a confiabilidad y validez de los resultados. Aplicados en forma correcta, los métodos estadísticos no permiten la demostración experimental de nada, pero sí sirven para asignar un nivel de confianza a un enunciado o para medir el error posible en una conclusión. La ventaja principal de los métodos estadísticos es que agregan objetividad al proceso de toma de decisiones. Las técnicas estadísticas, aunadas a una buena ingeniería o conocimiento del proceso y al sentido común, conducen por lo general a conclusiones sólidas.

Los métodos gráficos suelen ser útiles para esta etapa, en particular para presentar los resultados.

4.5. Normalización de valores

Por lo general, los estudios se realizan para recabar datos que posteriormente se emplearán para contrastarlos con otros valores, bien valores estándar bien resultados de otros experimentos. Sin embargo, antes de proceder a comparar los datos, es preciso transformarlos a una métrica común, mediante algún procedimiento de normalización conocido. Esta herramienta de homogeneización de valores, presenta a su vez importantes inconvenientes pues no existe un acuerdo generalizado en ámbitos profesionales o científicos sobre cuál es el procedimiento de normalización de valores más fiable o, al menos, el más adecuado (Cloquell, Santamarina et al., 2001). Esta falta de consenso ha propiciado que no exista un procedimiento rigurosamente dominante, lo cual tiene una consecuencia negativa directa sobre aquellos problemas en los que subyace una evaluación multicriterio.

Un inconveniente añadido de la normalización de valores y que posiblemente pudiera servir para explicar la falta de acuerdo en lo referente a su procedimiento, es que puede no afectar, en ocasiones, por igual a todas las alternativas por lo que exige una comprobación de la bondad de la misma en cada problema particular (Cloquell, Santamarina et al., 2001). De hecho, (Barba-Romero y Pomerol, 1997) afirman que "dicha normalización previa no es, por otra parte, neutral. En efecto, es perfectamente posible que el resultado final dependa del procedimiento de normalización de las evaluaciones", lo cual confirma la gravedad de la debilidad de los procedimientos empleados en la normalización.

En (Barba-Romero y Pomerol, 1997) se resumen los procedimientos de normalización que a juicio de los autores son los más destacables en la actualidad y que se han recogido en la Tabla 3:

	P_1	P_2	P_3	P_4
Definiciones	$v_i = \frac{a_i}{\max a_i}$	$v_i = \frac{a_i - \min a_i}{\max a_i - \min a_i}$	$v_i = \frac{a_i}{\sum_i a_i}$	$v_i = \frac{a_i}{\sqrt{\sum_i a_i^2}}$
Vector normalizado	$0<v_i\leq 1$	$0\leq v_i\leq 1$	$0<v_i<1$	$0<v_i<1$
Módulo de v	Variable	Variable	Variable	1
Conserva la proporcionalidad	Sí	No	Sí	Sí
Interpretación	% del máximo a_i	% del rango (max a_i – min a_i)	% del total $\sum_i a_i$	i-ésima componente del vector unitario

Tabla 3.- Principales métodos de normalización de vectores

El procedimiento P_1 es el más corriente, tiene una interpretación muy sencilla, como fracción de lo máximo posible, y respeta la proporcionalidad. El procedimiento P_2 es un perfeccionamiento del anterior, destinado a asegurar que las transformaciones cubrirán el intervalo [0, 1]. Si bien respeta la cardinalidad, no se garantiza que conserve la proporcionalidad. El procedimiento P_3 se utiliza frecuentemente, ofreciendo las mismas ventajas que P_1, aunque sin embargo da valores más pequeños y más concentrados. El P_4 es el procedimiento de normalización menos intuitivo. Ofrece la ventaja de permitir comparaciones de vectores de norma 1, pero al igual que el procedimiento anterior concentra valores.

Estos procedimientos de normalización precisan que los valores de los vectores de valores cumplan ciertos requisitos, entre ellos, que sean positivos, para lo cual podría solventarse con un cambio de origen previo, tal y como se ha visto en la sección 4.2.

4.6. Pruebas de contraste y medidas de asociación

Los tests o pruebas de contraste pueden clasificarse como *paramétricos* y *no-paramétricos*. Los **tests paramétricos** se basan en algún modelo que requiere que los datos tengan alguna distribución específica. En la

mayoría de los casos es asumible que algunos de los parámetros, los involucrados en un test paramétrico, estén distribuidos normalmente. Para comprobar la normalidad, por ejemplo, se puede emplear la prueba del Chi-cuadrado. Igualmente, los tests paramétricos precisan que los parámetros se puedan medir en al menos en un intervalo escalar. Cuando no se cumplen los requisitos del test paramétrico, éste no puede usarse. Los **tests no-paramétricos** no asumen nada sobre la distribución de sus parámetros, por lo que son más generales; luego, los tests no-paramétricos imponen menos restricciones a la hora de poder ser utilizados. Para decantarse por un tipo u otro de test, hay que tener en cuenta dos factores importantes: la *aplicabilidad* y la *potencia* propia de cada test. La aplicabilidad son las condiciones que deben cumplir los parámetros con respecto a su distribución y escala para que el procedimiento pueda se utilizado; de manera que un test sólo puede emplearse cuando las condiciones de aplicabilidad o bien se cumplen o bien son asumibles. La potencia de un test paramétrico es generalmente mayor que la de un no-paramétrico, interesando escoger el de mayor potencia.

(Wohlin, Runeson et al., 2000) describen con brevedad y claridad excepcional los tests más utilizados para el contraste de hipótesis, aportando un ejemplo ilustrativo para la ingeniería del software empírica. Aunque los tests no están descritos íntegramente, incluyen referencias bibliográficas alternativas donde se aborda el procedimiento en cuestión. A continuación se enuncian brevemente los tests o procedimientos estadísticos más frecuentemente empleados en contraste de hipótesis:

- **T-test:** Uno de los tests paramétricos más frecuentemente utilizados. El test se emplea para comparar las medias de dos muestras. Esto es, el diseño es un factor (una variable) con dos tratamientos.
- **Mann-Whitney**: Test no-paramétrico alternativo al t-test.
- **F-Test:** test paramétrico que se puede emplear para comparar las distribuciones de dos muestras.
- **T-test pareado**: un t-test para un diseño de comparación de pares.
- **Prueba de suma de rangos con signo de Wilcoxon para datos pareados**: Alternativa no-paramétrica para realizar un t-test pareado. Se emplea cuando se desea comparar la distribución de una variable en dos muestras de casos pareados, para ello trabajando con la diferencia entre las observaciones pareadas. Puede utilizarse para

probar o refutar la igualdad de las medias para observaciones pareadas de muestras dependientes.

- **Test de signos**: Alternativa no-paramétrica para realizar un t-test pareado, más simple que el test de Wilcoxon.
- **ANOVA (ANalysis Of VAriance)**: Una familia de test paramétricos que pueden ser empleados para comparar medias para al menos tres muestras independientes,.
- **Kruskal-Wallis**: Alternativa no-paramétrica para realizar una ANOVA**.**
- **Chi-cuadrado (χ^2)**: Una familia de test no-paramétricos que pueden emplearse cuando los datos están son frecuencias.
- **Test binomial**: Test no-paramétrico que puede emplearse para comparar las sucesiones de dos eventos complementarios.

La Tabla 4 muestra una ordenación de los tests atendiendo al tipo de diseño y a si son o no paramétricos:

Diseño	**Paramétrico**	**No-paramétrico**
Un factor, un tratamiento (= Una var. ; una muestra)		χ^2 Test binomial
Un factor, dos tratamientos, diseño totalmente aleatorio	t-test F-test	Mann-Whitney χ^2
Un factor, dos tratamientos, comparación pareada	t-test pareado	Wilcoxon Test de signos
Un factor, más de dos tratamientos	ANOVA	Kruskal-Wallis χ^2
Más de un factor	ANOVA	Friedman

Tabla 4.- Clasificación de los tests paramétricos/no-paramétricos para diseños distintos

Además de los tests de contraste mencionados, existe otro test no paramétrico denominado **test de Kappa-Fleiss** (***κ***). Este es el método más frecuentemente empleado *para evaluar el nivel de concordancia entre varios observadores.* Cohen introdujo en 1960 el índice o coeficiente kappa (Cohen, 1960), como una medición de la concordancia entre dos jueces donde cada uno califica un conjunto de objetos utilizando escalas nominales. La generalización del método a más de dos observadores fue extendida por Fleiss (Fleiss, 1971); y la incorporación de pesos se halla recogida en (Fleiss y Cohen, 1973). Kappa incorpora en su fórmula una corrección que excluye la concordancia atribuible al azar, permitiendo una estimación más precisa de la concordancia genuina. El índice kappa

puede tomar valores entre -1 y +1. Cuanto más cercano a +1, mayor es el grado de concordancia entre los observadores; por el contrario, cuanto más cercano a -1, mayor es el grado de discordancia entre los evaluadores. Un valor κ=0, refleja que la concordancia observada es precisamente la que se espera a causa exclusivamente del azar (López_de_Ullibarri y Pita, 1999). (Landis y Koch, 1977) propusieron unos márgenes para valorar orientativamente el grado de acuerdo en función del índice kappa, márgenes que se recogen en la Tabla 5. Por lo general, mientras que un valor inferior a 0,4 se considera "malo" (sin acuerdo, insignificante o bajo, según la interpretación de Landis y Koch), un valor superior a 0,75 indica una concordancia "excelente" según (Fleiss, 1981).

Kappa	**Grado de acuerdo**
<0	Sin acuerdo
0-0,2	insignificante
0,2-0,4	Bajo
0,4-0,6	Moderado
0,6-0,8	Bueno
0,8-1	Muy bueno

Tabla 5.- Interpretación de los valores del índice de kappa según el rango de valores de Landis y Koch

El índice kappa puede emplearse también cuando se quieren comparar dos o más métodos y no existe un estándar de referencia.

La realización de estudios de concordancia son habituales en investigación médica, en procesamiento de lenguaje natural y en ciencias conductuales y psicológicas. Muestra de ello son las siguientes referencias bibliográficas (Fleiss and Cohen 1973; Landis and Koch 1977; Siegel and Castellan 1988; Carletta 1996).

Capítulo 5
Estado del arte de los Sistemas con Bancos de Ítems Calibrados

No sólo la orientación práctica de las investigaciones realizadas en el ámbito de las teorías de test sino también un elevado número de sistemas computerizados de aprendizaje y/o de evaluación requieren la manipulación de grandes cantidades de ítems calibrados, y es en este último marco donde se pone de manifiesto su importancia y aplicación práctica.

Se considera que un banco de ítems calibrado es un conjunto de ítems o de preguntas referidas a un determinado dominio de aptitudes o conocimientos, almacenadas y clasificadas en función de su contenido y de sus propiedades métricas y cuyos parámetros se han estimado en una misma escala. Si bien este trabajo se centra en la calibración de ítems, hay que indicar que existen métodos para ayudar a confeccionar bancos de ítems, como son el recurrir a tests existentes en el mercado, tests empleados en centros docentes o, por el contrario, ser construidos ad hoc. Las personas interesadas en itemetría pueden ampliar sus conocimientos consultando, entre otras, las siguientes referencias (Cattell 1986; Prieto and Delgado 1996; Arruabarrena, Sanz-Santamaría et al. 2007).

En lo referente a la calibración de los ítems, no siendo un proceso excesivamente difícil, si que resulta un proceso extenso y costoso. Esencialmente existen dos alternativas: solicitar a expertos que efectúen

la calibración o bien emplear procedimientos estadísticos para realizar una estimación de los parámetros de los ítems.

En este capítulo se presenta la utilidad de los bancos de ítems seguida de sus áreas de aplicación, incluidos algunos sistemas de aprendizaje. Posteriormente se abordan las dos líneas de generación de calibraciones más extendidas, a partir de un conjunto de ítems ya construido: la calibración de ítems incorporando las aportaciones subjetivas de expertos y la calibración estadística de los ítems en el marco de la TRI. La primera línea es, hoy por hoy, la forma más extendida de acometer las calibraciones de ítems, si bien la documentación relativa a la misma es prácticamente inexistente. En el caso de la calibración estadística, por la que abogan los psicometras, la situación es diferente: la profusión documental en el ámbito teórico de las fases que constituyen el desarrollo de la calibración contrasta con la escasez documental relativa a estudios empíricos.

5.1. Utilidad de los bancos de ítems calibrados

Este apartado recoge una síntesis de las posibilidades y beneficios que pueden obtenerse de los bancos de ítems calibrados a la hora de construir tests y organizar el dominio de conocimiento de sistemas de aprendizaje (Arruabarrena, Sanz-Santamaría et al., 2007; Barbero, 1996):

- **Introducen flexibilidad en la evaluación** tanto en el campo psicológico como en el educativo. Las puntuaciones obtenidas por los usuarios evaluados en un conjunto cualquiera de ítems seleccionados proporcionarán una medida del rasgo en la misma escala. Esto posibilita la introducción de consideraciones prácticas de carácter específico en compilaciones posteriores de tests así como el desarrollo de sistemas personalizados de instrucción.
- **Posibilitan la construcción de test adaptativos informatizados** (TAIs). Si el banco de ítems está bien construido, permitirá compilar el test más adecuado de la longitud más conveniente para cada objetivo, evidentemente, siempre dentro de los límites marcados por el propio banco (Olea y Ponsoda, 2003). A su vez, los TAIs confieren

grandes ventajas que se han discutido en la literatura (véanse, por ejemplo, (Kingsbury y Weiss, 1983) y (Wainer, Dorans et al., 2000).

- **Facilita la comparación de resultados entre evaluados**, entre los ítems del banco y entre test (Wright y Bell, 1984). Teniendo en cuenta que cualquier test construido a partir de un banco de ítems y aplicado a una muestra de sujetos es automáticamente equiparado con los demás, para poder comparar los resultados obtenidos por una muestra de individuos no es preciso que todos ellos hayan respondido al mismo test. Se podrá seleccionar el conjunto de ítems que más se adecue al nivel de habilidad de cada individuo, ya que se dispone de la información necesaria para la compilación de tests que maximicen la información en los distintos niveles de habilidad (como, por ejemplo, las curvas características de los ítems, las funciones de información de los ítems o los tests).
- **Permiten reducir el tiempo de administración de los tests**. Teniendo en cuenta que se pueden comparar los resultados obtenidos por sujetos que han respondido a distintos ítems y que puede conocerse antes de su administración la relación entre los parámetros de cada ítem y el nivel de habilidad de los sujetos se puede disminuir considerablemente el tiempo de aplicación por sujeto y mejorar la precisión y rapidez en el proceso de compilación de tests.
- **Permiten pronosticar resultados de la administración de tests**. Cuando se dispone de un banco de ítems calibrado e igualmente se conoce de antemano el nivel de habilidad del individuo a ser evaluado, es posible predecir cuál será su comportamiento en un test antes de su administración. Dado que los ítems que componen el banco están debidamente calibrados, bastaría con calcular la probabilidad de que el individuo responda correctamente a cada uno de los ítems del test y, posteriormente, sumar dichas probabilidades.
- **Permiten realizar actualizaciones periódicas de las estimaciones de los parámetros de los ítems** a partir del uso eficiente de las respuestas de los evaluados a un conjunto de ítems, ya que cualquier conjunto de datos, aunque provengan de las respuestas de un único individuo a unos pocos ítems, puede ser incorporado en el sistema computerizado para la actualización sistematizada de las estimaciones.
- **Permiten la construcción de tests de gran calidad** puesto que los ítems incluidos en el banco son el resultado de un proceso de depuración a lo largo del cual se han eliminado aquellos que bien por

falta de ajuste al modelo elegido, bien por no adecuarse su contenido al banco, o por cualquier otro motivo, no fueran considerados pertinentes.

- **Ayudan a organizar y clasificar el material educativo** atendiendo a diferentes criterios establecidos (temática, grado de dificultad, nivel de profundización por áreas, etc.), propiciando un mayor aprovechamiento del material elaborado.

5.2. Ámbitos de aplicación de los bancos de ítems calibrados

Durante el periodo 1905-1908 los psicoanalistas franceses Alfred Binet y Théodore Simon desarrollaron una serie de procedimientos para **estimar la capacidad mental** de los sujetos a partir de la comparación de niños y adolescentes de diversas edades. Con los datos empíricos obtenidos al aplicar sus tests en poblaciones bien definidas, calibraron los ítems y definieron la escala de *edad mental* y a partir de ésta el *cociente de inteligencia* (Binet y Simon, 1905). Dichos expertos marcaron un antes y un después en la práctica de la medición psicológica, la de los *tests estandarizados*, y su pródigamente conocido trabajo fue la primera aplicación de bancos de ítems calibrados extensamente documentada en la historia.

Posteriormente, durante la I Guerra Mundial, y como consecuencia de la necesidad de **reclutar personas** para el ejército, tuvo lugar en EE.UU. la primera aplicación masiva de **test colectivos** - los Army Alfa. Pero el verdadero uso y aprovechamiento del potencial de los bancos de ítems se postergó hasta originarse el desarrollo computacional ya en los ochenta. Precisamente, a partir de esta década, el concepto de banco de ítems ha venido atrayendo la atención de agencias tanto públicas como privadas (Hiscox y Brzenzinski, 1980). Se han construido bancos calibrados en sectores tan diversos como en grandes organizaciones médicas, en las Fuerzas Armadas y en grandes compañías de tests.

No obstante, donde se ha observado un mayor desarrollo ha sido en el **campo educativo**, posiblemente, debido a la gran variedad de aplicaciones que permite dicho campo. De este modo, por ejemplo, pueden emplearse en el aula para construir tests que permitan a los docentes evaluar el nivel de conocimientos de sus alumnos (Nitko y Hsu, 1984; O'Brien y Hampilos, 1988) o bien pueden emplearse en las

distintas demarcaciones escolares a fin de informar a los centros y a la opinión pública acerca del rendimiento de los alumnos en distintas áreas curriculares (Douglas 1980; Hankins 1990; Moore 1994). Sin embargo, el área de aplicación de mayor calado está siendo su uso en diversos **estudios de evaluación del rendimiento académico con dimensión internacional**, como son los proyectos PISA, PIRLS, TIMSS, INES, entre otros muchos. Estos proyectos están mayormente auspiciados por la Organización para la Cooperación y el Desarrollo Económico (OCDE) y/o por la Asociación Internacional para la Evaluación de los Logros Educativos (IEA, del inglés Internacional Association for the Evaluation of Educational Achievement); y entre sus colaboradores dispares están los Ministerios de Educación de los respectivos países participantes en el estudio, agencias, asociaciones, consorcios, organismos e institutos tanto nacionales como internacionales entre los cuales caben destacar el Educational Testing Service (ETS), el Australian Council for Educational Research (ACER), el Japaneese Nacional Institute for Educational Research (NIER), el Dutch national institute for educational measurement (CITO), la Canada's national statistics agency (Statistics Canada), el Centro de Proceso de Datos de la IEA en Hamburgo (Alemania), el Instituto Nacional de Evaluación y Calidad del Sistema Educativo (INECSE, España) e incluso empresas, como la estadounidense WESTAT, entre otras.

En los proyectos de evaluación del rendimiento académico con dimensión internacional mencionados, además de una dirección transnacional, existe un Consejo de Países Participantes, que representa a todos los países, y determina las prioridades en materia de política educativa del proyecto en cuestión, a la vez que vigila el cumplimiento de las mismas. Se encarga de establecer las prioridades para el desarrollo de indicadores, para la preparación de los instrumentos de evaluación y para la presentación de los resultados. A su vez, los expertos de los países participantes colaboran también en grupos de trabajo encargados de actualizar los bancos de ítems sobre las diferentes áreas de conocimiento a evaluar, garantizando que los materiales renovados tengan cualidades de medición sólidas y que los instrumentos pongan énfasis en la autenticidad y validez educativa (Mullis, Martin et al., 2007a; OCDE, 2003). En la mayoría de este tipo de evaluaciones internacionales se utiliza la metodología de la TRI (MEC, 2007), lo que permite comparar resultados de alumnos a nivel nacional e internacional así como

confrontar resultados de evaluaciones anteriores, ello gracias a que no todos los ítems empleados en un mismo ciclo de evaluación son liberados (MEC, 2007; Mullis, Martin et al., 2007b) ya que se reservan como testigo para próximas evaluaciones. Tanto los análisis como los ítems liberados se recogen en informes nacionales e internacionales; y aunque en principio no son estudios académicos, porque están dirigidos sobre todo a los administradores y gestores de la educación, dichos informes pueden servir para orientar a los centros educativos y para detectar los puntos fuertes y débiles del currículo (Sjøberg, 2004), si bien las actividades y ejercicios utilizados en estas pruebas no suelen ser los mismos que se trabajan en los centros escolares. Los análisis transnacionales del rendimiento que logran los estudiantes permiten ampliar y enriquecer la visión nacional, proporcionando un contexto más amplio en el que interpretar los propios resultados.

La Tabla 6 recoge los proyectos internacionales para la evaluación del rendimiento del alumnado con más renombre. En la misma se pueden observar que existen proyectos que evalúan las mismas áreas pero a poblaciones con edades distintas, lo cual permite completar estudios y ver la tendencia de las poblaciones. Los ciclos o periodicidad de las réplicas de las evaluaciones de los proyectos varían entre 2 y 5 años. Igualmente el número de países participantes en unos proyectos y otros es distinto así como el volumen de la muestra evaluada en cada uno de los países; no obstante, cabe indicar que el volumen de la muestra es superior al 75% de la población en cada uno de los países y por proyecto.

En la tabla se han recogido las cifras de la última encuestación. Como se puede apreciar, el volumen es tal que los proyectos precisan necesariamente de bancos de ítems de considerable tamaño. Los bancos se renuevan en cada ciclo al liberarse parte de ellos. Sin embargo, es preciso mantener un bloque de los mismos (los ítems de anclaje) para realizar equiparación de puntuaciones entre administraciones consecutivas y hacer comparables los resultados nacionales e internacionales. Entre PISA, TIMSS y PIRLS, sin duda alguna, el proyecto PISA además de ser el más amplio, es el que más ha transcendido y mejor documentado está. De los desarrollos de las réplicas PISA 2000 y 2003 se ha podido acceder a información relativa a los procesos de construcción y calibración de ítems así como de equiparación de puntuaciones.

	OCDE/PISA[2]	**IEA/TIMSS**[3]	**IEA/PIRLS**[4]
Áreas de conocimiento evaluadas	Matemáticas, lectura, ciencias y solución de problemas	Matemáticas y ciencias	Competencia lectora
Población evaluada	15 años	9.5 y 13.5 años (4º Primaria y 2º ESO)	9.5 años (4º primaria)
Desde	1997	1991	2001
Ciclo del estudio	3 años	4 años	5 años
Última encuestación	2006	2007	2006
Países involucrados actualmente	Más de 50	Más de 60	45
Volumen examinados	[5.000-10.000] por país	En 1995: más de 500.000	[4.000-16.000] por país

Tabla 6.- Proyectos internacionales para la evaluación del rendimiento de alumnos

Relacionados con proyectos internacionales, y en la sociedad actual en la cual los indicadores de calidad y estadísticos internacionales están a la orden del día, va en aumento la difusión de **proyectos nacionales e internacionales** que precisan de bancos de ítems para compilar y administrar sus respectivos tests, para posteriormente **generar** los **valores de** sus propios **indicadores y proporcionar** informaciones sobre **tendencias** de los mismos. Así lo corroboran, entre otros, los estudios: *OCDE/INES*[5], *OCDE/PIAAC*[6], *OCDE/TALIS*[7], *IEA/TEDS-M*[8], *IEA/ICCS*[9], y *EBAFLS*[10].

---

[2] OCDE/PISA: Programme for International Student Assessment, (en castellano, Programa para la Evaluación Internacional de Alumnos), promovido por la OCDE. La web del proyecto es www.pisa.oecd.org. Referencias en castellano de este proyecto se pueden descargar directamente de la web del Instituto de Evaluación (www.institutodeevaluacion.mec.es) o bien del Ministerio de Educación y Ciencia (www.ince.mec.es/pub/) del cual depende el Instituto.

[3] IEA/TIMSS: Trends in International Mathematics and Science Study, (en castellano, Estudio Internacional de tendencias en Matemáticas y Ciencias) y está promovido por la IEA. La web del proyecto es timss.bc.edu/ o bien www.iea.nl. Referencias en castellano de este proyecto se pueden descargar directamente de la web del Instituto de Evaluación (www.ince.mec.es/pub/) o bien de (www.institutodeevaluacion.mec.es).

[4] IEA/PIRLS: Progress in Internacional Reading Literacy Study, (en castellano, Estudio Internacional de Progreso en Comprensión Lectora) y está promovido por la IEA. La web del proyecto es pirls.bc.edu/. Referencias en castellano de este proyecto se pueden descargar directamente de la web del Instituto de Evaluación (www.institutodeevaluacion.mec.es) o bien del (www.ince.mec.es/pub/).

5 OCDE/INES: International iNdicators of Education Systems.

6 OCDE/PIAAC: Programme for the International Assessment for Adult Competencies.

Para determinar el nivel lingüístico de los evaluados, existen otras muchas evaluaciones de características similares, que consisten en la administración de ítems elaborados por expertos a través de pruebas parejas y con criterios de corrección similares. Por ejemplo, y únicamente de nivel B1 según el Marco Común Europeo de Referencia para las Lenguas del Consejo de Europa[11] (Council_of_Europe, 2001) y para alumnos de 6º curso de educación primaria, están el Preliminary English Test (PET) gestionada por la Universidad de Cambridge, Lecteur de niveau seuil (Diplôme de Langue Francaise, niveau 2, (DELF B1 para abreviar)), el correspondiente diploma de alemán Zertifikat Deutsch, el Diploma Intermedio di Lingua Italiana (DILI), el Nivel B1 o Nivel intermedio de español (dentro de Diplomas de Español como Lengua Extranjera (DELE)) del Instituto Cervantes, el Certificat de Nivell Intermedi de Catalá (B1) organizado por el Departamento de Cultura de la Generalitat de Cataluña, la prueba B1 de euskera organizada por el Servicio de euskera de Gobierno Vasco, etc.

En el **ámbito de la educación universitaria**, la evaluación del rendimiento es parte consustancial de la enseñanza, puesto que tiene repercusiones de tipo legal al determinar aspectos tales como la adquisición de títulos académicos y de capacitación profesional. Por ello, es imprescindible que la evaluación se realice con garantías de igualdad, méritos y capacidad de los individuos evaluados. Así, cada vez son más las materias universitarias que disponen de amplios bancos de ítems para medir las destrezas y habilidades adquiridas por los alumnos universitarios (Caro, 1988; Martínez-Cervantes y Moreno-Rodríguez, 2002). Estos mismos planteamientos son aplicables también a pruebas de acceso a la universidad y a pruebas de destrezas de postgraduados como en los exámenes MIR (Médico Interno Residente) o PIR (Psicólogo

7 OCDE/TALIS: Teaching and Learning International Survey.

8 IEA/TEDS-M: Teacher EDucation Study in Mathematics.

9 IEA/ICCS: International Civic and Citizenship Study (en castellano; estudio internacional de la IEA sobre educación cívica y ciudadanía).

10 EBAFLS: European Bank of Anchor Items for Foreign Language Skills.

11 Marco Común Europeo de Referencia para las Lenguas del Consejo de Europa: es el documento base que se emplea para describir los niveles de aprendizaje de estudiantes de idiomas en Europa. Se trata de una iniciativa del Consejo de Europa con el fin de proporcionar un método de evaluar y enseñar que sea común para todas las lenguas de Europa.

Interno Residente), para lo cuales se disponen de extensos bancos de ítems. Por ejemplo, las pruebas de admisión a las universidades de suecia, las *SweSAT* (Swedish Scholastic Aptitude Test), se celebran dos veces al año, participando, entre ambas, cerca de 75000 personas y siendo el Swedish National Agency for Higher Education quien ostenta la coordinación nacional de dichas pruebas. En las mismas, los tests se vienen empleando desde hace 30 años siendo, hoy por hoy, la Teoría Clásica de los Tests (TCT) es el marco teórico subyacente de las pruebas, si bien existen ya trabajos para determinar si la TRI permitiría mejorar la calidad de las SweSAT. En concreto, y según (Stage, 2003), de mantenerse la SweSAT en su esquema actual, no se recomendaría el uso de la TRI en su confección, aunque si alguna versión futura la prueba se transformase en un TAI, entonces su uso sería forzoso.

Asimismo, y **fuera** ya **del ámbito de la educación reglada** y bancos de ítems transnacionales, existen bancos de ítems nacionales de gran envergadura para realizar **evaluaciones de oposiciones a gran escala**, destacando por el volumen de individuos que se someten a las mismas las oposiciones de distintos niveles para sanidad, administraciones públicas y docentes. Los ítems de dichas pruebas son redactados por expertos en el dominio del conocimiento, y las innumerables publicaciones actualizadas al respecto ponen de manifiesto la repercusión y alcance de dichas pruebas, como por ejemplo, (Aguilera y González, 2008; Tapias, 2008). En muchas ocasiones y con antelación a la celebración de la oposición son los mismos Ministerios, Institutos u organismos que organizan o gestionan las oposiciones quienes facilitan los ítems a preguntar, siendo los sindicatos quienes a continuación filtran las soluciones a los opositores.

Como se ha podido apreciar por lo expuesto hasta el momento, los bancos de ítems suscitan gran interés y además de los ya citados, se han desarrollado en innumerables países del mundo, incluyendo Estados Unidos (Burke, Kaufman et al., 1985; Henning, 1986), Australia (Cornish y Wines, 1977; Hill, 1985; Tognolini, 1982), Alemania (Weber, Kuhl et al., 2001), Austria (Kubinger, 1985), Escocia (Pollit, 1985), Holanda (Conejo, Guzmám et al., 2004; van Thiel y Zwarts, 1986), Inglaterra (Choppin, 1981; Elliot, 1983), España (Caro 1988; Barbero and Navas 1995; Pérez 2000; Conejo, Guzmám et al. 2004; Trella, Carmona et al. 2005), etc.

Nótese que en este apartado únicamente se han citado algunos de los bancos de ítems de mayor envergadura y de uso real. No obstante, desde primeros de este siglo, y gracias a los avances tecnológicos, los bancos de ítems no se hallan aislados, sino embebidos dentro de aplicaciones informáticas con proyecciones más amplias que albergar un mero banco de ítems, como son los sistemas de aprendizaje en su más amplio significado, abarcando así herramientas de autor, sistemas de evaluación basados en tests, redes bayesianas, STIs, SHAs, sistemas educativos inteligentes y adaptativos en red, etc. El desarrollo de estos tipos de sistemas está en plena expansión, y el lector interesado puede consultar diversas recopilaciones como las expuestas en (Kubeš, 2007; Weibelzahl, 2002) y (López-Cuadrado, 2008).

5.3. Proceso de calibración de ítems con expertos

Disponer de una batería o banco de ítems de una misma área de conocimiento no tiene mayor interés que tener agrupados un conjunto de ítems, en contraposición a que dichos ítems estén además calibrados cuyas ventajas y usos se acaban de exponer en los apartados precedentes. En cuanto a la propia construcción de los ítems, hoy por hoy lo habitual es que sean los especialistas o expertos en el área quienes los elaboren. Para ello se deben tener en cuenta cuestiones diversas como el objetivo del uso que se va a realizar de los ítems y el contenido y nivel de profundidad curricular que se quieren evaluar o enseñar a través de los ítems.

La calibración de ítems a partir de valoraciones subjetivas de expertos no es un proceso especialmente documentado, aunque sí el más frecuente para conseguir una calibración. Si bien existen publicaciones que abordan algunos aspectos teóricos concretos, apenas hay publicaciones que desvelen la labor desarrollada por los expertos involucrados. Por ejemplo, los ítems empleados en las pruebas a gran escala nacionales, transnacionales y oposiciones de la sección anterior, se efectúan con ítems construidos y calibrados, en primera instancia, por expertos. Posteriormente, una vez administradas las pruebas que contienen los ítems, en algunos casos, como en el estudio PISA y PIRLS, se recurre a la TRI para hacer comparables los resultados de distintas

ediciones nacionales e internacionales, para lo cual es preciso efectuar una calibración estadística de los parámetros de los ítems a partir de los datos recogidos de las propias pruebas. No obstante, el proceso concreto desarrollado por los expertos para construir calibraciones de ítems no se ha hallado documentado, encontrándose únicamente alguna referencia superficial sobre el mismo. Precisamente, dicha bibliografía específica es prácticamente inexistente, aunque existen salvedades como el banco de ítems del test de inteligencia de Binet-Simon a primeros del siglo pasado y el estudio PISA en la actualidad. En concreto, en PISA 2003 los ítems fueron elaborados y calibrados por comités de expertos nacionales y fue el equipo internacional de elaboración de pruebas del Consorcio de PISA quien negoció y decidió, por consenso, cuáles de los ítems propuestos desde los comités nacionales formarían parte de las pruebas internacionales y cuáles debían descartarse (OCDE, 2005). Sin embargo, no se ha hallado descrito en documento alguno el proceso concreto llevado a cabo por los comités nacionales para la generación de los ítems calibrados atendiendo a los parámetros de dificultad y competencia medida por los ítems (como, por ejemplo, en el área de las matemáticas las destrezas reflexión, conexiones, reproducción, etc.).

En (Helmer y Rescher, 1959), los autores argumentan que en aquellos ámbitos donde no es posible definir leyes científicas explícitas, el testimonio de los expertos es permisible como fuente de conocimiento científico. Por otro lado, es importante recalcar que métodos alternativos a no emplear expertos para manejar problemas de cierta enjundia pueden involucrar procesos prohibitivos en la práctica, en términos de costes y tiempo, de recolección y procesamiento de la información. Es así, como estas justificaciones aún son válidas para muchas aplicaciones cuando no se dispone de la información precisa, es muy costoso conseguirla o la evaluación requiere de datos subjetivos en los principales parámetros. No obstante, existe una creciente necesidad de incorporar información subjetiva (por ejemplo, análisis de riesgos) directamente en la evaluación de los modelos que tratan con problemas complejos a los que se enfrenta la sociedad, tales como, medio ambiente, salud, transporte, comunicaciones, sociología o educación. Esto ha dado lugar a multitud de sistemas informáticos que manipulan conocimiento gracias al uso, entre otros, de clases, taxonomías, ontologías, reglas (fuzzy), heurísticos y redes bayesianas. Este sin fin de aplicaciones, junto con sus publicaciones, corrobora la valía de la intervención de expertos y de la

incorporación de sus aportaciones y apreciaciones, en una medida u otra, en dichas aplicaciones. Por tanto, el problema se reduce a cómo obtener y utilizar dicho testimonio o, más específicamente, cómo combinar el testimonio de varios expertos en una declaración única.

En la bibliografía circundante a los sistemas informáticos en los que han participado individuos en calidad de expertos o especialistas en el dominio de conocimiento de la aplicación, y que se verán en las siguientes secciones, aparecen reiteradamente diversos aspectos de capital importancia. Estos aspectos acarrean tomas de decisión que, además de condicionar el desarrollo posterior de la investigación, proyecto o estudio a realizar, repercuten en los resultados finales. Los aspectos identificados se presentan agrupados en las siguientes tres subsecciones y son: aspectos relevantes relacionados con los propios expertos, con la planificación del proceso y con la validez de los resultados.

5.3.1. Sobre los propios expertos

Algunas publicaciones sobre desarrollos donde han colaborado expertos revelan el número de ellos que han participado y, en ocasiones, añaden además alguna indicación sobre la forma en la que estos han aportado su conocimiento. Sin embargo, otro tercer aspecto que es importante considerar es si puede, o no, suceder que los expertos abandonen su participación, y si así es, en qué medida puede suceder este hecho. A continuación se comentan más extensamente estos tres aspectos.

Para extraer conocimiento de una serie de personas se pueden emprender y combinar diversas alternativas. Algunas técnicas de evaluación o **herramientas para obtener información** son las expuestas en la sección 3.3, entre las que se encuentran los paneles de expertos, las entrevistas, las encuestas o cuestionarios, el benchmarking, los tests, los grupos de enfoque, los mapas conceptuales, el método Delphi, etc. Varias de estas técnicas son las que se emplean en evaluaciones formativas internas de sistemas informáticos con el objetivo de mejorarlos, por lo que la mayoría se pueden hallar más extensamente documentadas en libros dedicados a la evaluación de sistemas, por ejemplo (Harvey, 1998; Scriven, 1991) o también en libros de interacción persona computador, como por ejemplo (Dix, Finlay et al., 1998; Nielsen y Mack, 1994; Shneiderman, 1998). Aunque (Nielsen y Mack, 1994)

indican que las demostraciones de funcionamiento y uso a colaboradores y clientes pueden aportar un feedback interesante y provechoso a la hora de evaluar y mejorar los sistemas, en (Shneiderman, 1998) se puntualiza que las revisiones y participaciones formales de expertos han demostrado ser eficaces, como por ejemplo, en los estudios (Jeffries, Miller et al., 1991; Karat, Campbell et al., 1992).

El gran inconveniente de estos métodos es poder disponer, bien en plantilla, o bien como asesores externos, expertos o profesionales del área. Pero, ¿a cuántos expertos se debe consultar? En el caso de revisión de materiales, es sabido que diferentes expertos tienden a encontrar diferentes problemas, por lo que entre 3 y 5 revisores pueden resultar altamente productivos (Shneiderman, 1998). (Tessmer, 1993) aboga también por un número relativamente bajo de expertos, 2 ó 3 concretamente por cada área, aunque en la realidad se consulten únicamente 1 ó 2. En el ámbito de la evaluación experimental mediante experimentos controlados, (Dix, Finlay et al., 1998) recomiendan que el tamaño de la muestra sea lo suficientemente amplia y representativa con un mínimo de 10 sujetos. En cambio, diversos experimentos realizados por investigadores de la Rand Corporation (Dalkey, Brown et al., 1970), donde se empleaba el *método Delphi*, mostraron que la desviación media de las opiniones del grupo de expertos disminuye fuertemente con el número de participantes. De acuerdo con estos resultados, el **número de expertos que debieran participar** en un estudio de prospección es de siete. Estos autores apuntan que, si bien parece necesario para asegurar un buen funcionamiento del grupo un mínimo de 3 y 5 expertos, dicha cifra es algo dependiente del diseño del estudio; y habida cuenta de que el error disminuye notablemente por cada experto añadido hasta llegar a los 7 expertos, matizan que no es aconsejable recurrir a más de treinta expertos, pues la mejora en la previsión es muy pequeña y normalmente el incremento en coste y trabajo de investigación no compensa la mejora. Conforme con dichos experimentos, la validez de los resultados queda garantizada satisfactoriamente con un tamaño de panel superior a 13 expertos, aunque considerando la relación coste-beneficio, el número óptimo de expertos a involucrar es de 7.

Fijado el número de expertos a involucrar, otro factor importante a considerar es que hay que prever el porcentaje de los que iniciarán el proceso pero que no lo concluirán. Este hecho puede ocurrir principalmente en desarrollos en los que los expertos participan de forma

no remunerada. Resulta obligado, en consecuencia, comenzar el desarrollo con un número significantemente mayor de expertos que el que se ha establecido como adecuado. Luego, es preciso **prever** y considerar la **tasa de abandono de los expertos**.

5.3.2. Aspectos sobre la planificación

En esta sección se tratan diversas cuestiones a establecer durante la planificación del proceso de calibración de ítems.

Es incuestionable la importancia que tiene el realizar una planificación lo suficientemente exhaustiva, correcta y eficiente, no solo en los procesos de calibración de ítems sino también en todos aquellos procesos que consumen muchos recursos. La captación, gestión y control de los recursos humanos, como de los costes y de la información debe ser precisa y adecuada para alcanzar los objetivos propuestos. Existen innumerables libros que así lo corroboran, entre otros, y solamente en el área de la ingeniería del software están (Scriven 1991; Tessmer 1993; Worthen, Sanders et al. 1997), en el ámbito de la evaluación y en el ámbito de los experimentos empíricos controlados (Wohlin, Runeson et al., 2000). El Capítulo 3 y el Capítulo 4 han recogido sucintamente los fundamentos allí presentados y en los que se ha basado autora para abordar su trabajo.

Un aspecto recomendable a considerar durante la planificación es la **elaboración de una guía de trabajo** que proporcione a los participantes, y en particular a los expertos, aclaraciones sobre el estudio en el que participan, incluyendo **instrucciones y criterios comunes**. La guía ayudará a que los diversos individuos trabajen de forma coherente e integrada, puesto que el objetivo es construir una aportación única "común" a partir de aportaciones individuales. Ejemplo de ello son los *marcos de evaluación* en PISA (OCDE, 2003) o bien a nivel teórico las *guías* propuestas en (Worthen, Sanders et al., 1997) tanto para planificación como para conducción y ejecución de evaluaciones. No obstante, (Nielsen, 1993) matiza que es preciso llevar control del experimento pero sin que ello modifique en el criterio o la valoración del experimentado.

Además, es precisa la **revisión previa de los materiales** a administrar, para lo cual la supervisión por parte de revisores (Shneiderman, 1998) y la realización de **pruebas piloto** (Tessmer, 1993) pueden aportar mucha

información útil para detectar inconsistencias y hacer las correspondientes correcciones y ajustes allá donde sea preciso.

También en la fase de planificación es el momento de **concretar el análisis estadístico** que se efectuará a los datos muestrales. Deberá ser un análisis sensato, que considere el número de expertos participantes y el formato cualitativo o cuantitativo de los datos. El análisis de los datos se ha abordado más ampliamente en las secciones 3.3.3 y 4.4.4 de esta misma memoria. El análisis de los datos correspondiente a la calibración basada en la TRI se halla íntegramente en la siguiente sección.

5.3.3. La validez de los resultados

Para mitigar posteriores amenazas, es conveniente anticiparse y conocer cuáles son los aspectos que ponen en entredicho la valía de los resultados obtenidos en los estudios de investigación donde han participado expertos. Por ello, a la hora de planificar el desarrollo del proceso de calibración es conveniente sopesar aspectos relacionados con la valía de los expertos participantes y del proceso desarrollar, aspectos que se desarrollan en los siguientes párrafos.

En particular, dos problemas a menudo presentes en las críticas al uso de herramientas para obtener información de expertos son la calidad de una persona como experto, y el criterio de selección de la muestra de expertos.

La primera crítica, la de la calidad de una persona como experto, radica en la mayor o menor discrepancia que pudiera existir entre las valoraciones resultantes de la consulta a los expertos y las apreciaciones particulares. Esta discrepancia suscita desconfianza sobre el grado real de conocimiento atribuido a los expertos. Sin embargo, llegado el caso, se podrían tomar medidas al respecto. Así, y dependiendo de las características del estudio podría interesar **medir la precisión de los juicios de los expertos**, por ejemplo, empleando la técnica de evaluación "Metodología Pert y distribuciones beta" recogida en la sección 3.3.2, evidentemente siempre y cuando los requisitos para la aplicación de la técnica se cumplieran y tras sopesar la relación coste-beneficio que ello conllevaría. No obstante, y aunque no es habitual medir la bondad de los expertos, los más escépticos mantienen que los

juicios de expertos pueden ser mero reflejo de apreciaciones personales condicionadas (Worthen, Sanders et al., 1997).

Así mismo, la selección de la **muestra de expertos** es otro aspecto que va de la mano de la valía de estos. La dificultad de disponer de asesores expertos en el ámbito de la ingeniería del software es patente, y se ha tratado con anterioridad en la sección 4.4.2.2. Con el objeto de soslayar críticas sobre el criterio de selección de un subconjunto de expertos, una de las recomendaciones más aceptadas por la comunidad científica es aplicar el criterio de **máxima diversidad** definido por (Lang, 1995). Éste afirma que "la selección de un grupo tan diverso como sea posible minimiza el sesgo debido a la selección no aleatoria de los expertos".

En la valía de los resultados, otra cuestión clave es la fiabilidad de los procedimientos de medición empleados. Es necesario realizar un **análisis de la fiabilidad del proceso desarrollado**. Tradicionalmente, la variabilidad entre observadores se ha reconocido como una fuente importante de error en la medición (Fleiss, 1986; Landis y Koch, 1977). Consecuentemente, el objetivo de los estudios consiste en estimar el grado de dicha variabilidad. En este sentido, hay dos aspectos distintos que forman parte típicamente del estudio de fiabilidad: por una parte, el *sesgo entre observadores* – o bien dicho con menos rigor, la tendencia de un observador a dar consistentemente valores mayores que otro – y por otra parte, la *concordancia entre observadores (confiabilidad)* – es decir, hasta qué punto los observadores coinciden en su medición. Para realizar el contraste de confiabilidad, el método más frecuentemente empleado es el *índice Kappa-Fleiss* descrito en la sección 4.6. El análisis de fiabilidad de la calibración basada en expertos debiera incluir el estudio de ambos aspectos.

5.4. Proceso de calibración estadístico de ítems

En esta sección se exponen, primeramente, los fundamentos básicos de la TRI (Lord y Novick, 1968), incluidos supuestos y modelos más empleados y, seguidamente, las cuatro fases consecutivas en las cuales se puede desglosar la calibración estadística: se **administran los ítems a una amplia muestra de sujetos** (apartado 5.4.2) que sea representativa de la población, evitando con ello el sesgo en las estimaciones

posteriores; tras la recopilación de los resultados suele ser conveniente **realizar una serie de estudios** (apartado 5.4.3) que permitan detectar respuestas anómalas e ítems defectuosos y poder así retirarlos del proceso de calibración; a continuación, a partir de los datos depurados se efectuará **la estimación de los parámetros** de los ítems y las habilidades de los sujetos (apartado 5.4.4) de manera que se maximice el grado de ajuste entre los datos y el modelo; justamente, en la fase posterior es preciso **verificar el ajuste de los datos empíricos al modelo** de la TRI utilizado (apartado 5.4.5).

5.4.1. Fundamentos de la Teoría de Respuesta al Ítem (TRI)

La TRI constituye un nuevo enfoque en psicometría para la medición de variables psicológicas y educativas, que permite superar algunas de las deficiencias de la TCT, y cuyas peculiaridades proporcionan un modelo teórico excelente para la elaboración de test adaptativos computerizados. Si bien el origen de estos modelos puede encontrarse en (Lazarfeld, 1950), dada la complejidad de los cálculos para su aplicación, únicamente empezó a difundirse y utilizarse gracias a programas de computación específicos. La bibliografía relacionada con la TRI es muy abundante, por lo que este apartado tan sólo se limita a ofrecer una rápida exposición de los conceptos fundamentales y características principales de la misma. Las personas interesadas en ampliar conocimientos pueden consultar, entre otros, los libros (Hambleton y Swaminathan, 1985; Hambleton, Swaminathan et al., 1991; Lord, 1980; Muñiz, 1996).

La TRI intenta dar una fundamentación probabilística al problema de la medición de rasgos inobservables. La TRI asume que existe una función matemática que relaciona la competencia de los sujetos con la probabilidad de que estos respondan correctamente a los ítems. En otras palabras, que dada la competencia de un sujeto en la variable medida, cuando a éste se le presenta un ítem se conoce la probabilidad que tiene de acertarlo.

De acuerdo con (Hambleton, Swaminathan et al., 1991) son varías las características principales de la TRI como alternativa a la TCT. Partiendo del hecho de que son modelos centrados en el ítem más que en el test, hay que indicar que las características de los ítems son individuales y se

establecen de manera independiente, sin depender del grupo del cuál fueron obtenidas. Igualmente, los modelos permiten obtener estimaciones de la habilidad de los evaluados, que son independientes del conjunto específico de ítems que se les haya administrado, e incluso permiten determinar la precisión con la que cada individuo es medido. Existen algunas otras ventajas de la TRI que explican su popularidad, pero la más importante para fines prácticos, es que los examinados no necesitan contestar el mismo conjunto de ítems a fin de ser comparados con un misma escala (Ozen y Reise, 1994).

5.4.1.1. Supuestos de la TRI

Como se ha mencionado anteriormente, la TRI proporciona una familia de modelos matemáticos que se sustentan en el cumplimiento de diversos requisitos teóricos que deben cumplir los datos y que se asumen los verifican. En este apartado se enuncian los supuestos o condiciones que se deben verificar cuando se usa la TRI como modelo de examen de comportamiento (Hambleton, Swaminathan et al., 1991; Weiss y Yoes, 1991).

El primer supuesto, aparentemente trivial, se hace tanto en la TRI como en la TCT y señala que **si el administrado conoce la respuesta a un ítem, entonces probablemente lo responderá correctamente**. A veces esta suposición se refrasea en negativo, es decir, que si el examinado ha respondido incorrectamente a un ítem del test, entonces probablemente no conocía la respuesta correcta de dicho ítem.

El supuesto de **unidimensionalidad** señala que los ítems sólo sirven para medir un rasgo. Aunque en realidad la TRI proporciona modelos que admiten la posibilidad de que la respuesta de un ítem sea atribuible a más de una habilidad, lo cierto es que en la mayoría de los casos se supone que se evalúa sólo una. Así, todos los ítems de un banco miden la misma variable (conocimiento, habilidad, actitud o rasgo de la personalidad). La mayoría de los tests (y, por lo tanto, los ítems incluidos) que se usan en la actualidad están diseñados para medir una única habilidad, de manera que la asunción de la unidimensionalidad no es excesivamente restrictiva, si bien hay que verificarla.

En la TRI se asume que el comportamiento de un examinado con un nivel de conocimiento estimado y ante un ítem *i* puede predecirse y modelarse estadísticamente mediante una función monótona creciente

denominada la **Curva Característica del Ítem** (CCI) (Tucker, 1946) que representa gráficamente la relación no lineal entre la habilidad θ del examinado (eje horizontal) y la probabilidad *P(θ)* de que éste responda correctamente al ítem (eje vertical) (Figura 10). Esta función especifica que, a medida que el nivel del rasgo aumenta, la probabilidad de responder correctamente también aumenta. Luego, la CCI expresa gráficamente la probabilidad de que un individuo con cierto nivel de conocimiento responda correctamente al ítem.

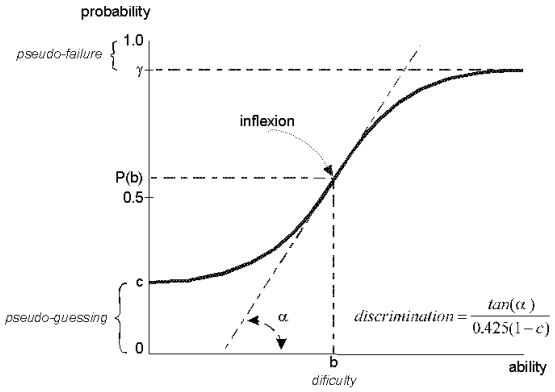


Figura 10.- CCI para un modelo logístico de 4 parámetros

El principio de invarianza es una de las características más relevantes de la TRI, y se enuncia en dos sentidos. La propiedad de **invarianza con respecto al grupo de administrados** dice que los parámetros de un ítem son propios del ítem, constantes, invariables e independientes de la habilidad de los sujetos que lo respondan. Invariablemente, en la TRI las propiedades de los tests se miden únicamente en función de las características de los ítems que lo componen. A su vez, la **invarianza con respecto al conjunto de ítems administrados** señala que la habilidad del examinado es constante, invariable durante la administración del test (aunque puede variar a lo largo de la vida del administrado), e independiente de los ítems que se utilicen para estimarla.

Por último, y no por ello menos importante, está el supuesto de **independencia local**, el cual está muy ligado con el principio de invarianza y el supuesto de unidimensionalidad, y expresa que la probabilidad de un examinado de contestar correctamente un ítem no depende de las respuestas dadas a los otros ítems del test. Técnicamente, esto significa que no existe una correlación entre los ítems para individuos con el mismo nivel de habilidad (independencia estadística). Este axioma, crucial para la TRI porque los ítems se combinan basándose en ella, se viola si, por ejemplo, el contenido de un ítem en el

test da pistas para conseguir la respuesta correcta de un ítem posterior. Como consecuencia, si se cumple este supuesto se puede asumir que el orden en que se administran los ítems dentro del test es irrelevante en lo concerniente a los resultados (Wainer y Mislevy, 1990).

Llegados a este punto, hay que recordar que los modelos de TRI son modelos "fuertes", ya que los supuestos pueden resultar difíciles de confirmar por los datos del test. Sin embargo, aun pudiendo ser complicada la verificación de los supuestos, es indispensable que exista un ajuste entre el modelo y los datos del test que sean de interés, para que la teoría subyacente al test permita hacer distintas inferencias a partir de las puntuaciones obtenidas en el mismo por los evaluados.

5.4.1.2. Modelos de la TRI

Existen diversos modelos de la TRI, pero todos tienen en común el uso de alguna función matemática para especificar la relación entre el comportamiento del examinado (factor observable) en un test y los rasgos o habilidades latentes (factores no observables) que se supone están implícitas en el desempeño del test.

Los modelos se pueden clasificar en los denominados modelos **paramétricos**, aquellos en los que las CCIs están caracterizadas por funciones conocidas o bien en los denominados modelos **no paramétricos o de ojiva normal**, aquellos en los que las CCIs se obtienen directamente de resultados estadísticos asemejándose a funciones de distribución normal acumulada. En la categoría de los modelos paramétricos, hay diversas alternativas para caracterizar las CCIs, siendo entre ellas la más ampliamente utilizada la familia de curvas logísticas de uno, dos o tres parámetros (1PL, 2PL o 3PL) (Birnbaum, 1968) que se definen de la siguiente manera:

$$P(u_i = 1 \mid \theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-1.7a_i(\theta - b_i)}}$$

Ecuación 2.- Modelo logístico de 3 parámetros

donde u_i=1 indica que el sujeto ha respondido correctamente el ítem *i*; y con $P(u_i=0 \mid \theta)=1- P(u_i=1 \mid \theta)$ en caso contrario. θ es el nivel de habilidad del examinado, es decir, lo que se mide mediante el test y habitualmente toma un valor real del intervalo [-3, 3]. Luego, $P(u_i=1 \mid \theta)$ es la probabilidad de que un examinado elegido al azar con habilidad

θ conteste correctamente al ítem *i*. Según el modelo, cabe indicar que los individuos con bajas habilidades (<0.0) tienen escasa probabilidad, en contraposición a los que tienen habilidades muy elevadas. Los tres parámetros que determinan la figura de la curva son:

- *Parámetro* a_i. Se denomina **índice de discriminación**. Cuanto mayor sea este valor, mayor capacidad tendrá el ítem para decidir si la habilidad del evaluado corresponde a un nivel superior o a uno inferior con respecto a la dificultad del ítem.
- *Parámetro* b_i. Es el **índice de dificultad del ítem**. Indica la dificultad del ítem. La dificultad se define utilizando la misma escala que para la habilidad del evaluado: El rango de valores de este parámetro es a lo largo del eje horizontal (-∞,+∞), si bien a efectos prácticos es el rango (-3, 3), siendo el 0 su punto central. Cuanto mayor sea el valor de este parámetro, más difícil será el ítem.
- *Parámetro* c_i. Se denomina **pseudoacierto** y expresa la probabilidad de acertar un ítem cuando se desconoce la respuesta correcta, es decir, la probabilidad de acertar al azar.

La Ecuación 2 se emplea para modelar el modelo 3PL. El modelo de dos parámetros asume que el factor de pseudoacierto es cero (c_i=0), mientras que en el modelo de un parámetro se supone que la discriminación del ítem es la misma para todos los ítems (a_i=1). Elegida una determinada función matemática para la CCI, según los tres parámetros citados tomen unos valores u otros, las curvas adoptarán distintas formas.

Los modelos anteriores son modelos **dicotómicos** en los que se establece para cada ítem sólo dos posibles respuestas (correcta e incorrecta). Son los modelos más utilizados y en los cuales se centra este estudio.

5.4.2. Administración de los ítems

Para calibrar estadísticamente un banco de ítems el primer paso consiste en administrar los ítems a una muestra aleatoria, representativa y al menos de varios cientos de sujetos, a tenor de los trabajos empíricos publicados. Para mayor detalle pude consultarse (López-Cuadrado, 2008).

A cada sujeto se le puede o bien administrar todo el banco de ítems o bien se puede fragmentar éste en bloques o **subtests** y administrar un único subtest a cada individuo. No obstante, cuando el tamaño del banco de ítems es de considerable tamaño, la aplicación completa del banco a un único individuo puede resultar inaceptable o inviable, aun siendo ello posible. En estos casos está ampliamente extendido el decantarse por la administración fraccionada de los ítems del banco. Sin embargo, esta opción añade una nueva problemática al poder tener cada subtest su propia métrica, como consecuencia de haber sido administrado a un conjunto de individuos distinto. En esta ocasión, la aproximación habitual para proyectar los resultados de cada subtest a una métrica común, que será la métrica del banco de ítems final, es el **diseño de anclaje de los ítems** y sobre el cual versa la siguiente sección (5.4.2.1). En términos generales, el diseño de anclaje consiste en incluir en cada subtest un número de ítems comunes, de manera que toda la muestra evaluada los responda; y con las estimaciones de estos ítems, denominados **ítems de anclaje**, se efectuará posteriormente la conversión de cada una de las escalas métricas a una única escala (Navas, 1996). A este último proceso se le denomina **equiparación de puntuaciones** y se hablará del mismo en la sección 5.4.2.2. Obsérvese, en realidad la equiparación de puntuaciones se efectúa cronológicamente más adelante, cuando se disponga de las estimaciones para los parámetros de los ítems de cada subtest, y sin embargo, se presenta en el contexto de la administración de ítems ya que se origina como consecuencia de aplicar únicamente parte de los ítems a cada individuo y del diseño de anclaje.

5.4.2.1. Diseño de anclaje de ítems

Básicamente el diseño de anclaje consiste en, a la hora de repartir los ítems en varios subtests, hacer que algunos ítems estén incluidos en más de un subtest simultáneamente, de manera que suceda **solapamiento de ítems** entre subtest. Existen diversas alternativas en cuanto a la forma de diseñar el solapamiento entre los tests al igual que existe disparidad de opiniones sobre el número de ítems que deben de componer el conjunto de ítems de anclaje o de solapamiento.

Cuando se trata de calibrar ítems dicotómicos y se pretende emplear el modelo logístico de tres parámetros, el anclaje más adecuado es el que

sigue un **diseño de grupos no equivalentes de ítems comunes** (Kolen y Brennan, 1995), siendo además el esquema más utilizado (Muraki, Hombo et al., 2000). Y entre las alternativas de modos de solapamiento de los ítems de anclaje, la variante que produce resultados más consistentes es aquella en que *todos los susbtests incluyen el mismo conjunto de ítems de anclaje*.

En el diseño de grupos no equivalente de ítems comunes los tests, que comparten el mismo conjunto de ítems de anclaje, se administran a grupos de individuos cuyas puntuaciones se espera difieran en cada muestra y que sus habilidades se distribuyan a lo largo de toda la escala de dificultad. No obstante, con el fin de minimizar diferencias, es conveniente que la muestra con la cual se efectúa la calibración sea lo más parecida posible a los individuos que en un futuro responderán los ítems ya calibrados. Cuanto mayores sean las diferencias entre las muestras de los sujetos, más dificultades tendrán los procesos estadísticos involucrados en la posterior equiparación de las puntuaciones. (Kolen y Brennan, 1995) indican que mientras que diferencias medias del 10% en la desviación típica de los ítems comunes se pueden aceptar, si éstas llegan al 30% posiblemente, y dependiendo de la técnica del proceso de equiparación de puntuaciones que se emplee, posteriormente darán lugar a problemas.

A la hora de elaborar los subtests, es importante que los ítems de anclaje sean una muestra representativa del banco total y que a su vez ofrezcan buenas expectativas (esto es, ítems que a priori se espera, tengan alta discriminación y bajo pseudoacierto). Los ítems de anclaje deberán aparecer en todos los subtests en la misma posición (Eignor, 1985; Muraki, Hombo et al., 2000) y los subtests deberán tener la misma longitud sin llegar a fatigar al administrado (Renom y Doval, 1999). Es igualmente recomendable realizar las administraciones de los tests de calibración del modo más similar posible a las futuras administraciones (Wainer y Mislevy, 2000).

En lo referente al número de ítems que deben componer el conjunto de anclaje, desde un punto de vista estadístico, cuántos más ítems de anclaje se utilicen, menor será el error durante la fase de equiparación de puntuaciones (Kolen y Brennan, 1995). Así, llevado al extremo, cuando el conjunto de anclaje lo constituyen todos los ítems del banco, el error de equiparación será nulo ya que únicamente habría un test y por

consiguiente una escala métrica; si por el contrario, y en pro de tests más cortos, se pueden emplear conjuntos de anclaje de unos pocos ítems pero en detrimento de la precisión de las puntuaciones. Intentando alcanzar un equilibrio entre el error por equiparación y los tamaños del conjunto de anclaje y de los tests, lo más habitual es que para subtests con más de 40 ítems el tamaño del conjunto de anclaje sea al menos una quinta parte de la longitud del test, salvo que el test sea muy largo, en cuyo caso bastaría con 20 ó 30 ítems de anclaje (Kolen y Brennan, 1995; Navas, 1996). En la misma tónica, y en cuanto a número de administrados, (Vale, Maurelli et al., 1981) sugiere administrar al menos a 30 personas los tests con 15 ó 25 ítems comunes.

5.4.2.2. Equiparación de puntuaciones

La equiparación de puntuaciones es un proceso estadístico que permite unificar las puntuaciones de los distintos subtests, y cuyas dificultades probablemente serán distintas, con el fin de poder compararlas en una escala de habilidad con origen y unidad comunes. Técnicamente, cuando se ha utilizado un diseño de anclaje para la administración de los ítems, se dirá que estos están *calibrados* una vez se haya efectuado la equiparación de sus parámetros, hasta entonces, los parámetros únicamente estarán *estimados*. Para poder realizar una equiparación correcta de los subtests deben satisfacerse los requerimientos de simetría, equidad, invarianza entre grupos y unidimensionalidad de los ítems (Angoff, 1984; Hambleton y Swaminathan, 1985; López-Cuadrado, 2008).

Para controlar la arbitrariedad de la escala de medida, lo habitual en la TRI es fijar la media ($\mu=0$) y la desviación típica ($\sigma=1$) de las estimaciones de habilidad y dificultad (Ogasawara, 2001; Santisteban y Alvarado, 2001). A partir de aquí, si un modelo se ajusta a un conjunto de datos, entonces cualquier transformación lineal que se aplique a la escala de habilidades también lo hará, siempre y cuando también se transformen los parámetros de los ítems (Kolen y Brennan, 1995). Así pues, una vez estimados los parámetros del modelo, y después de haber desechado aquellos ítems que comprometen la calidad del banco, es posible unificar la métrica de los parámetros de los ítems. Para ello, basta con transformar cada una de las puntuaciones de habilidad obtenidas en los diferentes subtests para disponer de una única escala común a todo el

banco (Hambleton y Swaminathan, 1985). Gracias a que la escala es única, las puntuaciones obtenidas en cualesquiera tests formados a partir del banco calibrado serán directamente comparables.

Cuando se trabaja con el **modelo logístico de tres parámetros**, (Hambleton y Swaminathan, 1985) indica que la transformación lineal debe realizarse según las siguientes ecuaciones siendo α la *pendiente de la recta* y β la *ordenada en el origen*.

$$a'=a/\alpha \qquad b' =\beta+ \alpha b \qquad \theta'=\beta+ \alpha\theta$$

Ecuación 3.- Transformación del parámetro de discriminación, de dificultad y de habilidad

Nótese que los valores de pseudoacierto no se ven alterados como consecuencia de un cambio de escala, de ahí que no se defina una ecuación de transformación para el parámetro c. Para que la relación entre los parámetros de los ítems y las habilidades medidas (o lo que es lo mismo, la curva característica del ítem) no se vean alteradas, es preciso realizar simultáneamente las tres conversiones, sean cualesquiera los valores de α y β utilizados, obteniéndose:

$$P(\theta') = c + \frac{1-c}{1+e^{-Da'(\theta'-b')}} = c + \frac{1-c}{1+e^{-Da(\theta-b)}} = P(\theta)$$

Ecuación 4.- Igualdad entre la CCI antes y después de la transformación

Existen diversas técnicas para establecer valores de α y β, siendo las más empleadas, por su simplicidad de cálculo, los *métodos basados en momentos media-sigma* y *media-media*. Sin embargo, hay que indicar que existen otros métodos de cálculo, no existiendo aún consenso a la hora de decantarse por uno u otro (Navas, 1996). Concretamente, el **método media-sigma** *o método estándar de la media y la desviación típica* (Marco, 1977) proporciona los valores de α y β a partir de las medias ($\overline{b_1}$ y $\overline{b_2}$) y desviaciones típicas (s_{b1} y s_{b2}) de las estimaciones del parámetro de dificultad de los ítems de anclaje en cada subtest (S_1 y S_2) tal y como indica la ecuación siguiente:

$$\alpha = \frac{s_{b1}}{s_{b2}} \qquad \beta = \overline{b_1} - \frac{s_{b1}}{s_{b2}}\overline{b_2}$$

Ecuación 5.- Valores de α y β según el método media-sigma

En cambio, el método *media-media* (Loyd y Hoover, 1980) recurre a la media de la estimación en cada subtest de los parámetros de discriminación de los ítems de anclaje ($\overline{a_1}$ y $\overline{a_2}$) para obtener el valor de

α, y posteriormente hace lo propio con las medias de las estimaciones de dificultad ($\overline{b_1}$ y $\overline{b_2}$) para despejar β mediante el siguiente par de ecuaciones:

$$\alpha = \frac{\overline{a_1}}{\overline{a_2}} \qquad \beta = \overline{b_1} - \frac{\overline{a_1}}{\overline{a_2}}\overline{b_2}$$

Ecuación 6.- Valores de α y β según el método media-media

Durante el proceso de equiparación pueden identificarse dos tipos de errores: el *error aleatorio*, relacionado con la muestra de sujetos, y el *error sistemático* de equiparación. El primero se puede reducir incrementando el tamaño de la muestra. Por su parte, el segundo puede considerarse como un sesgo originado normalmente como consecuencia de la violación de algún de los supuestos o condiciones de la metodología empleada (Kolen y Brennan, 1995; Navas, 1996). El método más habitual para estimar el error total de una equiparación consiste en definir un índice, como por ejemplo el *error cuadrático medio*, que considera los errores de todas las puntuaciones y puede descomponerse en dos términos, uno asociado al error aleatorio y otro ligado al error sistemático.

5.4.3. Depuración de los datos

Una vez administrados los tests y registrados sus resultados, y antes de proceder a la estimación de los parámetros, es recomendable efectuar algunos análisis con el fin de detectar y depurar anomalías en los mismos. A la hora de analizar las matrices de respuesta, (Renom y Doval, 1999) proponen realizar tres tipos de acciones: *filtrado de la obtención y captura de datos*, *análisis convencionales* de cada subtest para detectar ítems incompatibles con los modelos de la TRI, y *verificación de las pautas de respuestas* de los examinados. Los dos últimos análisis deben realizarse de forma combinada y cíclica para que la calibración posterior pueda sea satisfactoria.

El análisis de la *unidimensionalidad del banco de ítems* igualmente puede realizarse con antelación a la estimación de los parámetros. Y aunque dicho análisis propiamente pertenece a la fase de verificación del ajuste al modelo de la TRI (véase 5.4.5), en la práctica suele adelantarse al no ser necesario conocer de antemano los valores de los parámetros para efectuarlo.

Como consecuencia de los análisis realizados, puede suceder que sea necesario retirar ítems del banco. Por ejemplo, durante el proceso de calibración y tras la fase de depuración de datos en el sistema eCat se eliminaron el 10 % de los ítems (Olea, Ponsoda et al., 1996) y en el del TAI de ingreso a Hezinet el 19% (López-Cuadrado, Pérez et al., in press).

5.4.3.1. Filtrado de la obtención y captura de datos

Para que un sistema funcione correctamente, la entrada al mismo no debe incluir datos incoherentes o incorrecciones. Las anomalías pueden tener distinto origen: pueden haber sido generadas por los evaluados (dobles opciones o marcas, opciones fuera de rango, omisiones, etc.), el sistema de registro de respuestas puede producir matrices de respuestas incorrectas (tabulación incorrecta, doble respuesta, etc.) si incumple el protocolo de almacenamiento de las mismas, o incluso, las anomalías pueden producirse durante la propia transmisión de los ficheros. En este sentido, lo más conveniente es registrar las pautas de respuestas en bruto para poder posteriormente detectar ítems defectuosos, y corregir los datos brutos para poder realizar un posterior análisis exploratorio y conocer las distribuciones de las puntuaciones.

5.4.3.2. Análisis clásico de ítems

Mediante este estudio se pretende descartar ítems con características extremadamente desfavorables. El análisis hace especial hincapié en el análisis de incompatibilidades de los ítems de anclaje, ya que estos repercuten directamente en la equiparación de puntuaciones. Por ello, y antes de proceder a la estimación de los parámetros, suele realizarse un análisis clásico (en términos de la TCT) de los ítems para cada uno de los subtests completos y para el conjunto de ítems de anclaje.

El **análisis de frecuencias de las respuestas seleccionadas** es uno de los estudios que se realiza, el cual permite detectar determinados errores en el diseño de los ítems, como son enunciados confusos, distractores que nunca se seleccionan o ítems que se aciertan siempre o tienen elevadas probabilidades de acierto por conjetura o eliminación de alternativas. Las tasas de acierto y error dan una idea de la dificultad del ítem, pero además es conveniente analizar los perfiles de acierto y de

omisión, esto es, la tendencia a no responder los ítems con relación a la habilidad de los evaluados.

La **discriminación clásica de los ítems** (en términos de la TCT) es otro aspecto que se explora, esto es, la relación existente entre la puntuación de ítem y la puntuación total del subtest. En un estudio de correlación ítem-total lo esperable es que los sujetos fallen un ítem tiendan a conseguir una puntuación baja, mientras que los que lo acierten tiendan a obtener una puntuación alta. Así, lo habitual es eliminar del banco aquellos ítems que no alcance un índice mínimo de correlación de 0.3. El análisis de medias, desviaciones y curtosis permite observar cómo se distribuyen las respuestas sobre las categorías.

Para denotar el grado de correlación ítem-total, el más utilizado es el índice de consistencia interna alfa de Cronbach (Cronbach, 1951). Este índice se basa en la variabilidad de los ítems y es un indicador de discriminación, homogeneidad y consistencia interna, sirviendo como estimador de la fiabilidad del test. Si en un conjunto de ítems se detectan relaciones consistentes –el valor de la alfa de Cronbach se aproxima a 1-, se puede más o menos acertadamente suponer que dicho conjunto es homogéneo y, por tanto, unidimensional. Sin embargo, para estudiar la unidimensionalidad del banco de ítems se suele recurrir a otros tipos de análisis, como es el factorial exploratorio.

5.4.3.3. Verificación de las pautas de respuestas

El objetivo de este análisis es descartar evaluados (es decir, todas las respuestas emitidas por un mismo sujeto) como consecuencia de haber detectado en sus respuestas patrones absurdos (siempre escogen la misma opción u omiten casi todas las respuestas, etc.). En este punto del proceso de calibración únicamente se pueden descartar algunos evaluados por pautas absurdas o incoherentes en su actuación, ya que la detección exhaustiva de patrones aberrantes de respuesta no se podrá completar hasta más adelante, cuando se hayan obtenido las estimaciones de los parámetros de los ítems y las habilidades de los sujetos evaluados. En dicho momento se podrán describir otras conductas más sutiles, que debieran descartarse, y que pueden provocar serios problemas a la hora de ajustar los resultados a un modelo de respuesta al ítem o bien alterar los resultados de la calibración. Por ejemplo, debieran descartarse las respuestas de los sujetos que aciertan lo difícil y fallan lo fácil, copian a

otro durante la administración de los subtests, responden al azar, o se equivocan en la colocación de las respuestas sobre el papel. De ello se hablará en la sección 5.4.5.

5.4.4. Estimación de los parámetros

Tras haber realizado los análisis recién mencionados, los datos se hallan depurados y se habrán retirado eventualmente algunos ítems (de anclaje o no) del banco. Llegados a este punto, es necesario determinar el modelo de la TRI que se va a utilizar para definir las características psicométricas del banco de ítems, dado que el número de parámetros difiere siendo unos métodos de estimación más adecuados que otros.

Por cuestiones de economía y facilidad de manejo, los primeros bancos de ítems fueron calibrados según el modelo de Rasch (esto es, de 1 parámetro), que resulta de gran utilidad cuando los parámetros han de estimarse a partir de una muestra pequeña de datos (Lord, 1980). Para ítems de respuesta abierta, el modelo logístico más adecuado es el de dos parámetros. No obstante, lo habitual es disponer de bancos de ítems de respuesta cerrada, y más concretamente de elección múltiple; por lo que en la actualidad se recomienda y tiende a utilizarse el modelo logístico de tres parámetros (Weiss 1983; Millman and Arter 1984; Hambleton, Zaal et al. 1991; McBride 1997). En lo referente a los modelos multidimensionales, estos imponen muchas más restricciones (Wise y Kingsbury, 2000), están en fase de experimentación (Wainer, Dorans et al., 1990), requieren muestras mucho mayores y demandan gran cantidad de especificaciones (Hontangas, Ponsoda et al., 2000). De entre los modelos unidimensionales, el logístico de tres parámetros es el más recomendable al ser el más general y es además fácilmente reducible al de dos o al de un parámetros, aunque también tiene hoy por hoy sus detractores (Renom, 1993). Para ítems de respuesta múltiple dicotómicos, el modelo de tres parámetros es el más adecuado, algo en lo que coinciden la mayoría de los autores (Santisteban y Alvarado, 2001), dado que se esperan diferentes índices de discriminación y probabilidades elevadas de acierto al azar, más aún cuando no se permite la alternativa de omitir ítems (Wainer y Mislevy, 2000).

Establecido el modelo de la TRI a utilizar, existen diversos procedimientos para determinar los valores de los parámetros del modelo. En concreto, es posible emplear la **técnica de máxima**

verosimilitud condicionada para realizar estimaciones conjuntas consistentes de habilidades y parámetros (Andersen, 1970); sin embargo, este método puede resultar computacionalmente tan costo que sea imposible en algunas situaciones (Wainer y Mislevy, 1990), y no sirve para los modelos logísticos de dos y tres parámetros.

Otro tipo de procedimiento es la **estimación máximo verosímil conjunta** (Birnbaum, 1968). Dicho procedimiento consiste en elegir como valores para los parámetros del modelo aquellos que maximizan la probabilidad de que se produzcan las respuestas obtenidas en la aplicación de los subtests, o lo que es lo mismo, los que hacen más admisibles los resultados. La lógica de esta estimación trata de encontrar los valores de los vectores θ, a, b y c que conjuntamente maximicen la función de verosimilitud L, cuya expresión, una vez aplicado el logaritmo neperiano a ambos lados de la igualdad, viene dada por la Ecuación 7 (Baker, 1992; López Pina, 1995). Sin embargo, la estimación conjunta de habilidades y parámetros plantea dos inconvenientes: por un lado, exige la eliminación de las puntuaciones extremas – todo aciertos o todo fallos- tanto para ítems como para sujetos (Santisteban y Alvarado, 2001), y por otro, conforme aumenta el tamaño de la muestra aumentan también el número de parámetros y habilidades a estimar. Y lo que es peor, no está demostrado que según se incrementa el número de ítems y de sujetos los estimadores converjan hacia sus valores reales. De hecho, se han encontrado conjuntos de datos para los que el método de máxima verosimilitud conjunta falla. A raíz de lo comentado, no es de extrañar que muchos autores prefieran no utilizar este procedimiento (Molenaar, 1995).

$$\ln L(u \mid \theta, a, b, c) = \sum_{i=1}^{N} \sum_{j=1}^{n} \left[\; u_{ij} \ln P_{ij} + (1 - u_{ij}) \ln(1 - P_{ij}) \right]$$

Ecuación 7.- Función de verosimilitud conjunta

El método de **estimación máximo verosímil marginal** evita los problemas que surgen en el método de la estimación de la máxima verosimilitud conjunta. A diferencia de la estimación anterior, el procedimiento de máxima verosimilitud marginal proporciona **consistencia** a la estimación de los parámetros, y es independiente del tamaño de la muestra; de manera que aunque se incremente la cantidad de examinados, no será necesario estimar más parámetros.

La primera implementación destacable de estimación de máxima verosimilitud fue propuesta por (Bock y Lieberman, 1970) y puesto que realizaba numerosos y costosos cálculos sólo resultaba viable para tests de no más de 12 ítems (Baker, 1992; Hambleton y Swaminathan, 1985). Y pese a que probablemente es la técnica más utilizada (Glas, 2000), no está exenta de problemas. Por ejemplo, sus resultados pueden divergir cuando las matrices de respuesta son poco densas, concretamente si a cada sujeto se le ha administrado muy pocos ítems o si, pese a ser muchos, estos no se distribuyen abarcando todo el rango de la escala de habilidades. Además, puede volverse inestable cuando se viola alguno de los postulados básicos de la TRI, como la independencia local (Chen y Thissen, 1997). La variante *máximo verosímil marginal de puntuaciones acumuladas* (Chen y Thissen, 1999) puede funcionar mejor en estas circunstancias, si bien produce algo más de sesgo en sus estimaciones.

El procedimiento máximo verosímil marginal puede generalizarse de varias formas (Glas, 2000). Así, (Swaminathan, Hambleton et al., 2003) manifiestan que la **estimación bayesiana marginal modal** de (Mislevy, 1986) surgida para evitar tener que estimar simultáneamente los parámetros es una generalización de la **estimación conjunta bayesiana** de (Hambleton y Swaminathan, 1985) que mejora los resultados. Si bien al ser un método bayesiano resulta complejo, a su vez ofrece la ventaja de que la estimación es **directa**, no haciendo falta imponer ningún tipo de restricción. Además, incorporar información previa acerca de los parámetros de los ítems permite reducir el tamaño de la muestra y mejorar los resultados de la calibración (Swaminathan, Hambleton et al., 2003).

Gracias a lo avances informáticos, hoy en día existen diversos programas informáticos que efectúan automáticamente la estimación de parámetros según alguno de los procedimientos recién presentados, y que pueden consultarse en (García Cueto, 1996).

5.4.5. Ajuste de los datos al modelo

Tras administrar los ítems y depurar los datos registrados, se procede a la estimación de los parámetros. En este punto del proceso de calibración, y para garantizar que los valores de los parámetros de los ítems son independientes de la muestra de los sujetos a partir de la que han sido estimados, es preciso **verificar que el modelo elegido se ajusta a las**

respuestas obtenidas durante la fase de administración y, a su vez, **comprobar que se cumplen las restricciones** impuestas por el mismo.

Estudiar el ajuste de las estimaciones al modelo de la TRI (esto es, comprobar que sus valores se corresponden con los resultados obtenidos empíricamente durante la administración de los subtest) es necesario para dotar de valor psicométrico a los ítems (López Pina, 1995). Si el modelo utilizado y el banco de ítems con el que se va a trabajar no están ajustados, las propiedades de la TRI carecen de validez: la información que se tenga de los ítems no será fiable y, consecuentemente, tampoco las estimaciones de habilidad que de ellos devengan.

Los modelos de la TRI fundamentan su flexibilidad en la realización de suposiciones muy restrictivas, tanto que prácticamente ningún conjunto de datos las satisfarán jamás (Lord y Novick, 1968). Por ello, el objetivo será intentar ajustar los datos al modelo más general, por basarse en suposiciones menos restrictivas. Además, se puede emplear el *test de la razón de verosimilitud generalizado*, que permite comparar las estimaciones para dos modelos, tales que uno es la generalización del otro; de manera que se pueda decidir si ambos se ajustan igualmente al mismo conjunto de datos, y en consecuencia pueda uno quedarse con el más simple y restrictivo.

Como consecuencia de la fase de ajuste al modelo, puede suceder que se retiren algunos ítems del banco por no respetar alguno de los supuestos. De hecho, cuanto antes se retiren los ítems contraproducentes para la evaluación tanto mejor. Por ejemplo, el efecto de los ítems defectuosos es especialmente crítico en los TAI debido a su relativamente corto tamaño. Por ello, se suele dedicar considerable esfuerzo a la hora de garantizar la calidad de los ítems del banco (Hambleton, Zaal et al., 1991). Aparte de los ítems retirados en la fase de depuración de los datos, se pueden descartar aquellos que no superan el análisis de unidimensionalidad, y una vez estimados los parámetros, pueden igualmente retirarse, por ejemplo, los ítems que tienen discriminación negativa o excesivamente pequeña, los que ofrecen valores altos de pseudoacierto, los que son extremadamente fáciles o difíciles, etc..

El estudio de la **bondad de ajuste** al modelo se efectúa a distintos niveles (López Pina, 1995). Por una parte, se verifica el ajuste de cada ítem (o de todo el banco como unidad) al modelo: cuanto más se

acerquen las puntuaciones estimadas por las curvas características de los ítems a las puntuaciones empíricas extraídas de los subtests, mejor será el ajuste al modelo; por otra parte, también se verifican el cumplimiento de los supuestos de la TRI y las características esperadas del modelo, tales como la invarianza de los parámetros de los ítems. Al igual que para la estimación de los parámetros de los ítems, existen programas informáticos que efectúan automáticamente la bondad del ajuste.

5.4.5.1. Bondad de ajuste de los parámetros de los ítems

El análisis de la bondad del ajuste consiste en observar la correlación existente entre los datos obtenidos tras la administración de los subtests y los pronosticados a partir de las estimaciones de los parámetros. Existen varios métodos para analizar estadísticamente el grado de bondad de ajuste, destacando los *basados en la distribución Chi-cuadrado* (χ^2) y el *análisis de residuales*. Sin embargo, hay que puntualizar que los índices basados en χ^2 deben interpretarse con mucha precaución, ya que son sensibles al tamaño muestral, y el análisis de residuales es aún una técnica sin una base estadística que la sustente, pero que parece configurarse como la más prometedora después de los estudios realizados por Hambleton y colaboradores (López Pina y Hidalgo Montesinos, 1996). (Hambleton y Swaminathan, 1985) proponen completar cualquiera de estos dos estudios con una comparación gráfica entre la media predicha por el modelo y la observada.

Los **análisis basados en la distribución** Chi cuadrado (χ^2) (Traub y Lam, 1985) se derivan directamente de la función de ajuste máximo verosímil, y tratan de contrastar la hipótesis nula de que las curvas características observadas y las pronosticadas para cada ítem o subtest no difieren demasiado. Para ello, dividen la escala de habilidad en varios pequeños intervalos, y calculan, para cada uno de ellos, las proporciones empírica y esperada de respuestas correctas de los sujetos cuya habilidad queda dentro del intervalo. A partir de estos dos valores se construye un determinado *estadístico de ajuste* con el que se contrasta, para cada ítem, la hipótesis nula de que las proporciones de respuesta encontradas en cada intervalo (CCI observadas o empíricas) no son diferentes de las esperadas según el modelo de respuesta al ítem (CCI pronosticadas o teóricas), o lo que viene a ser lo mismo, que el ítem se ajusta al modelo.

Si en el intervalo de confianza establecido no puede rechazarse la hipótesis nula, se considera que el modelo ajusta a los datos.

Por su parte, en **el análisis de residuales** (Hambleton, Swaminathan et al., 1991) el continuo de habilidades también se divide en varios intervalos. El procedimiento consiste en construir la distribución de frecuencias de los residuales obtenidos tras el ajuste del modelo a los datos empíricos (residuales reales) y la de los obtenidos al ajustarlos a una matriz de datos generada, mediante simulación, a partir de las estimaciones de los parámetros de los ítems y de las habilidades de los sujetos obtenidas en la fase anterior (residuales simulados). Así, para cada ítem *j* e intervalo *k*, se define el *residual* (r_{jk}) como la diferencia entre la proporción de acierto teórica o esperada según el modelo (P_{jk}) y la empírica (Pe_{jk}). Suponiendo que cada intervalo sigue una distribución normal $N(0,1)$, puede definirse, a partir de r_{jk}, el *residual estandarizado* (z_{jk}), una medida que considera el error de muestreo asociado con la proporción de respuestas correctas, y se expresa matemáticamente mediante la Ecuación 8 (Hambleton, Swaminathan et al., 1991; Muñiz, 1997).

$$z_{jk} = \frac{P_{jk} - Pe_{jk}}{\sqrt{Pe_{jk}(1-Pe_{jk})/N_k}}$$

Ecuación 8.- Residual estandarizado

Una vez comparados los residuales estandarizados en cada intervalo para cada uno de los modelos de respuesta al ítem, podrá decidirse cuál de ellos explica mejor el comportamiento de los ítems. Por ejemplo, si para todos los ítems e intervalos se verifica que $-1.96 < z_{jk} < 1.96$, entonces puede decirse, con una confianza del 95%, que existe un ajuste de los ítems al modelo. En general, cuanto más se alejen de cero los valores de los residuos, peor será el ajuste al modelo.

5.4.5.2. Restricciones y características esperadas del modelo

La comprobación del ajuste de los modelos no debe sustentarse únicamente sobre un estudio estadístico individual de cada uno de los ítems, sino que es preciso poner a prueba los supuestos que fundamentan el modelo especificado. En este sentido, es preciso probar que se cumplen los supuestos de *unidimensionalidad* e *independencia local de*

los ítems, junto con la propiedad de *invarianza de los parámetros*. La invarianza debe ser probada, y no asumida en base a la utilización de la TRI, ya que sobre esta propiedad recae la justificación de utilizar estos modelos y no otros.

Para comprobar empíricamente que los ítems del banco satisfacen el supuesto de unidimensionalidad, resulta necesario verificar el análisis de unidimensionalidad separadamente a cada subtest completo y, también, al conjunto de ítems de anclaje. El procedimiento más habitual para los modelos de respuesta al ítem unidimensionales y dicotómicos es el **análisis factorial exploratorio** (Lumsden, 1976) sobre la matriz de correlaciones entre los ítems. El análisis factorial exploratorio consiste en formular un modelo lineal que asocia las variables observadas (en este caso, las respuestas de los examinados a los ítems) y las dimensiones o factores que en ellas influyen. Esta relación se suele establecer en base al número de rasgos subyacentes, que para el análisis de unidimensionalidad interesa que sea sólo uno, donde todos los demás valen cero, esto es, que no tengan efecto sobre las observaciones. Una vez se obtiene una factorización de las diferentes dimensiones que evalúan los ítems, se podrá determinar hasta qué punto la estructura de covariación entre los ítems puede resumirse en un único factor. Puesto que para el conjunto de anclaje se dispone de una muestra muchísimo mayor, lo habitual suele ser realizar, adicionalmente, un **análisis factorial confirmatorio** sobre la matriz de correlaciones entre los ítems de anclaje. A diferencia del análisis factorial exploratorio, que como su nombre indica sirve para descubrir o explorar los factores que explican los patrones de respuesta de los examinados, en este caso se trata de confirmar la hipótesis de que sólo existe una dimensión o factor relacionado con dichas variables. Tanto si se trata del análisis exploratorio como del confirmatorio, (Santisteban y Alvarado, 2001) no aconsejan emplear ni el índice de *correlación de Pearson* ni el índice de *correlación phi*, ya que pueden dar falsos positivos. En su lugar, lo recomendable y habitual es utilizar la **matriz de correlaciones tetracóricas**, propuestas para medir la relación entre variables dicotómicas cuyas variables continuas subyacentes se asume que siguen una distribución normal (López Pina y Hidalgo Montesinos, 1996). Una vez extraídas las raíces latentes de la matriz de correlaciones tetracóricas entre los ítems, y teniendo en la diagonal las comunalidades o proporciones de varianza explicadas por los diferentes factores, la existencia de un primer factor notablemente superior al segundo, que a

su vez apenas difiera del resto, es condición suficiente para que se dé la unidimensionalidad del banco de ítems (Lord, 1980). Sea cual sea el método utilizado par determinar la unidimensionalidad del banco (otros métodos se pueden consultar en (Cuesta, 1996)), generalmente se realizan reiteradamente varios análisis factoriales, de manera que tras cada análisis se retiran los ítems que merman la unidimensionalidad antes de repetir el proceso.

Una vez probado que los ítems del banco verifican el principio de unidimensionalidad, podrá decirse que también se satisface la **propiedad de independencia local**, algo que no ocurre a la inversa (Renom, 1993). No obstante, para estudiar específicamente el supuesto de la independencia local hay que evaluar la matriz de varianzas y covarianzas, o la de correlaciones de las puntuaciones de los sujetos, fijando varios intervalos según el nivel de habilidad, y asumiendo el cumplimiento de este postulado cuando los valores de los elementos fuera de la diagonal principal son próximos a cero (Santisteban y Alvarado, 2001).

Igualmente hay que verificar que los ítems del banco satisfacen los **principios de invarianza** lo que equivale a realizar dos comprobaciones: que las estimaciones de habilidad se obtienen en la misma escala con independencia del conjunto de ítems administrados y que los parámetros de los ítems del banco no dependen de la muestra empleada durante el proceso de estimación. Aunque existen distintas formas para evaluar los supuestos de invarianza, lo habitual es comparar los parámetros de los ítems y los valores de habilidad estimados conjuntamente en la fase anterior del proceso de calibración, pudiéndose realizar tantas comprobaciones dividiendo en dos o más partes el banco de ítems (por ejemplo, ítems fáciles y difíciles, identificadores pares e impares, etc.) o la muestra de sujetos (por ejemplo, hombres y mujeres, aleatoriamente, etc.). Dado que los principios de invarianza suponen que las estimaciones de los parámetros y de las habilidades deben ser iguales con independencia de la muestra que se emplea para calibrarlos, si no ha habido errores de muestreo y el modelo de respuesta al ítem se ajusta correctamente a los datos, tanto los parámetros de los ítems como los valores de habilidad debieran ser iguales para cada grupo. Sin embargo, los resultados empíricos rara vez serán iguales. El coeficiente de correlación de Pearson entre las diferentes estimaciones puede dar una idea de lo parecidos que son los valores de habilidad estimados por los diferentes tests.

PARTE 3
Evaluación de calibraciones mediante TRI y expertos

*La **Parte 3** describe el trabajo realizado para la comparación de la calibración de ítems con expertos frente a la que se obtiene mediante los métodos de calibración TRI. La evaluación se centra en dos aspectos: primero, en la propia estimación de la dificultad de los ítems obtenida y, segundo, en los recursos consumidos. Para ello se ha tomado un banco de ítems utilizado para la ubicación del nivel inicial de los nuevos alumnos de Hezinet y se han confeccionado dos calibraciones, cada una basada en uno de los métodos antes mencionados.*

*El **capítulo 6** presenta la calibración de ítems empleando valoraciones subjetivas de expertos. La herramienta elegida son cuestionarios en formato papel y lápiz en la que se pide a expertos que los respondan como si de un alumno nuevo se tratasen, que informen sobre la dificultad de los ítems objeto de estudio, además de una valoración sobre las destrezas que evalúa el ítem. Para recoger la muestra se han seguido dos métodos distintos que han supuesto niveles de abandono diferentes por parte de los expertos, pero que también han tenido un consumo de recursos diferente por parte de los participantes activos.*

*El **capítulo 7** presenta la calibración de ítems utilizando el modelo logístico de 3 parámetros de la TRI (3PL-TRI). Se ha optado por realizar un proceso de calibración off-line puro porque sólo así se puede garantizar la calidad psicométrica. Se ha dividido el banco de ítems en varios subtests con un diseño de anclaje mediante grupos no equivalentes con ítems comunes. La administración de los ítems se ha hecho mediante cuestionarios electrónicos de forma supervisada y no supervisada con unos criterios muy estrictos para la validación de cada una de las administraciones para asegurar su validez. El trabajo de estimación de los parámetros de los ítems se ha hecho utilizando una métrica común.*

*El **capítulo 8** presenta la evaluación multicriterio realizada entre las calibraciones de los ítems por un lado, atendiendo a los valores de dificultad estimados, y por otro, atendiendo a los costes asociados a su producción. Mientras que el contraste entre las estimaciones de dificultad busca determinar si existe o no diferencias estadísticamente significativas entre las calibraciones realizadas, el contraste de costes busca determinar la viabilidad y eficiencia de los desarrollos alternativos.*

*El **capítulo 9** presenta una propuesta de proceso de negocio para realizar cualquiera de las dos calibraciones estudiadas a partir de los resultados de los experimentos controlados y análisis realizados.*

Capítulo 6
Calibración de ítems con Expertos (CE)

Al proceso de calibración siguiendo las valoraciones de expertos se le ha denominado, para abreviar, CE. Para llevarlo a cabo, fue necesario conseguir la valoración de una serie de expertos en lengua vasca. Esta tarea se realizó con sendos experimentos desarrollados mediante pruebas de campo en las que se entrevistó a expertos utilizando cuestionarios y que se llevaron a cabo de forma secuencial. Se ha denominado a cada una de las pruebas como Prueba con Expertos 1 (abreviado PE1) y Prueba con Expertos 2 (PE2).

Los tamaños de las muestras de las pruebas PE1 y PE2 se establecieron inicialmente en 10 valoraciones por ítem y por tipo de prueba, superando las recomendaciones coste-beneficio de (Dalkey, Brown et al., 1970) quienes recomiendan siete expertos. En el supuesto de que para algún ítem específico el número de valoraciones obtenidas y aceptadas fuera inferior a 5, el ítem no se calibraría, ya que (Shneiderman, 1998) indica que al menos son necesarias entre de 3-5 opiniones dependiendo del tipo del estudio.

En las siguientes secciones se presentan cada una de las pruebas realizadas.

6.1. Prueba con expertos 1 (PE1)

El objetivo de la prueba era la recopilación de 10 valoraciones de expertos por cada uno de los 252 ítems del banco de partida. Con la sobrecaptación de valoraciones se buscaba asegurar que, una vez

depurada la muestra, por cada ítem se alcanzasen las 7 valoraciones recomendadas.

Se crearon 8 cuestionarios de 42 preguntas cada uno, con 12 ítems de anclaje por si hubiera necesidad de equiparar las valoraciones. En el anexo A2 se puede ver uno de estos cuestionarios. A continuación se detallan los aspectos más importantes del experimento desarrollado.

6.1.1. Participantes

Los sujetos participantes activos fueron un coordinador y ejecutor principal, un supervisor general del proceso y un encargado de transcribir la información recopilada a una base de datos así como de redactar diversos informes. Además, se distinguieron dos tipos de sujetos pasivos: los *revisores* y los *expertos*, que colaboraron de forma voluntaria y no remunerada, ya que el grupo de investigación no disponía de recursos económicos.

Los **revisores** fueron 5 filólogos o lingüistas de euskera que se encargaron de hacer las pruebas piloto, donde se revisaron los cuestionarios atendiendo a aspectos como su duración o corrección sintáctica y gramatical.

Los **expertos** fueron 95 profesores de euskera de 22 euskaltegis que se encargaron de responder los cuestionarios que se les presentaron. Cada euskaltegi se comprometió a rellenar entre 3 y 5 cuestionarios.

6.1.2. Metodología

El método seguido para desarrollar la prueba de campo se describe en los siguientes párrafos:

Primero **se diseñó** y se preparó el material en **cuestionarios**. Como el volumen del banco de ítems era considerable y puesto que la colaboración de los sujetos pasivos era voluntaria, era materialmente imposible que cada uno de ellos valorase todos los ítems. Por ello hubo que fraccionar el banco en varios cuestionarios (específicamente en 8 en la PE1) y se decidieron los datos sobre los

ítems a recoger a través de dichos cuestionarios, tal y como se describe en el apartado 6.1.3.

Para **captar los revisores** se contactó directamente con miembros del entorno del equipo de investigación y de euskaltegis que cumplían con el perfil requerido.

Se realizaron **pruebas piloto** para detectar errores en los cuestionarios tal y como se especifican en el apartado 6.1.4.

Los **expertos se captaron** a partir de una lista de euskaltegis del País Vasco ordenada por cercanía al puesto de trabajo de la desarrolladora principal, ya que la entrega y recogida de cuestionarios se haría en persona, facilitando así la resolución de dudas y reforzando el compromiso de los sujetos. Para presentar el cuestionario se remitió una carta electrónica indicando el objetivo del estudio e invitando a tomar parte en él. Se acordó con aquellos centros que estuvieron dispuestos a colaborar una fecha de entrega y recogida de los cuestionarios junto con el número de cuestionarios que podían completar. Para cumplir con el objetivo inicial, fue necesario estimar cuál sería la tasa de abandono de los expertos. En este caso se estimó en un máximo de un 20%. Esto significó que para conseguir el objetivo era necesario comprometer un 120% de cuestionarios si la estimación era acertada, de ahí que para recuperar 80 cuestionarios completados por otros tantos expertos, se acordara la realización de 95.

La **administración de los cuestionarios** entre los diversos euskaltegis se hizo de forma escalonada, haciendo la entrega y recogida de los cuestionarios en mano a un responsable de cada centro, en quien se delegó la distribución y recogida de los mismos una vez estuvieran completados.

Como se comentará en una sección posterior, se pretendía llevar un **control** exhaustivo **de los costes temporales y económicos invertidos** durante todo el ciclo de vida de la prueba de campo. Se hizo uso de la técnica de calidad PDCA para **identificar aspectos mejorables** de los subprocesos.

6.1.3. Diseño de los cuestionarios

Se decidió que el tiempo necesario para responder a los mismos fuera de 45 minutos aproximadamente, restricción que afectaba al número de ítems incluidos en la sección correspondiente. La estructura de los cuestionarios estaba formada por una *introducción*, seguida de una *recogida de datos personales*, los *ítems a valorar* y, las *aportaciones propias* que pudiera manifestar el experto.

La **introducción** presentaba el objetivo del trabajo y las instrucciones de rellenado del cuestionario ilustradas con ejemplos.

La **recogida de datos personales** del participante obtenía algunos datos del experto con fines estadísticos como la edad, sexo, titulación superior, titulación lingüística o experiencia laboral en el área.

La parte central del cuestionario estaba dedicada a presentar un **subconjunto de los ítems a valorar**. Se solicitó que el experto proporcionase la respuesta correcta del ítem, además de la *destreza gramatical* y *nivel de dificultad*.

El experto tenía que elegir entre las **destrezas** consideradas por los pedagogos que colaboraron durante el desarrollo de Hezinet: *comprensión de textos, comprensión oral, conectivas, declinación, expresión escrita, ortografía, sintaxis, sufijos, verbos y vocabulario.* Además, se añadió una nueva destreza denominada *otra* por si el experto pudiera precisar una que no estuviera en la lista.

Para medir el **nivel de dificultad**, se optó por una escala de tipo Likert con los doce niveles en los que dividía el conocimiento de euskera la institución HABE en sus currículos (HABE, 1999). El primero de ellos correspondía al nivel inicial siendo el 12 el más avanzado.

Finalmente, se incluía el apartado de **aportaciones propias**. Este era una pregunta abierta para recabar comentarios o sugerencias por parte de los participantes.

6.1.4. Pruebas piloto

Las pruebas piloto se desarrollaron con **5 revisores**: cuatro eran filólogos vascos, que trabajaban o habían trabajado en desarrollos y

estudios lingüísticos del euskera en la Universidad del País Vasco, habían realizado traducciones de textos (libros incluidos) y algunos eran, además, profesores de asignaturas de euskera técnico de la UPV/EHU.

Durante las pruebas piloto, los revisores detectaron algún detalle a corregir en los cuestionarios y 265 errores individuales en 184 ítems, de los que **189 fueron erratas intranscendentes** cuya corrección resultó automática. En general se trataba de cuestiones relativas al estilo de puntuación, a normas académicas ortográficas o de algún otro desliz al transcribir los enunciados.

También identificaron un total de **76 errores graves** repartidos en 56 ítems. Entre estos casos había contextos en los que era posible que más de una opción de respuesta fuera correcta y otros en los que todas las opciones eran incorrectas, violaciones de alguna regla gramatical, ambigüedades o imprecisiones en el enunciado. En estos casos, se solucionaron 36 problemas. No obstante, 22 de estos ítems, pese a haber sido corregidos, quedaron **marcados como potencialmente erróneos**, dado que no se podía garantizar que los cambios fueran a ser eficaces. Los identificadores de los ítems marcados fueron 1, 13, 16, 59, 77, 102, 135, 146, 148, 152, 170, 178, 186, 191, 198, 202, 225, 229, 237, 240, 242 y 249. Un informe más detallado de los problemas detectados y su resolución se puede encontrar en (Arruabarrena, 2005).

Además, estas pruebas sirvieron para determinar si la estimación realizada del número de ítems por cuestionario se ajustaba al objetivo de 45 minutos que se había planteado. En cuanto a la *estimación temporal* cabe indicar que los revisores que no añadieron apenas comentarios emplearon algo más de media hora. Los que hicieron comentarios muy exhaustivos sobrepasaron la hora, mientras que el resto rondaron los 45 minutos.

6.1.5. Resultados

La administración de la primera prueba con expertos se desarrolló desde febrero de 2004 hasta noviembre de 2004. Durante el desarrollo se invitó a un total de 32 euskaltegis a participar en el

estudio. De estos, 22 centros accedieron y acordaron completar 95 cuestionarios por otros tantos expertos.

MUESTRA	Consultados	Acordados	Reales
N. expertos		95	74 (78%)
N. de centros (euskaltegis)	32	22	20 (91%)
Media de expertos x centro		4.3	3.7 .

Tabla 7.- Tamaño de la PE1

Sin embargo, como puede comprobarse en la Tabla 7, durante el desarrollo de la prueba, el 22% (100-78%) de los expertos abandonó, por lo que finalmente se recogieron 74 cuestionarios de 20 centros (el 91% de los centros acordados). De media se entregaron 4.3 cuestionarios por centro y se recogieron 3.7. En términos porcentuales, tal y como refleja el gráfico de la Figura 11, el 9% de los centros abandonó su participación en la prueba, el 27% rebajó el nivel de sus compromisos completando menos cuestionarios que los acordados y el 64% cumplió sus compromisos.

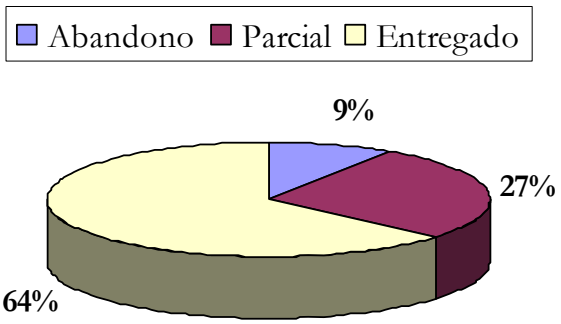


Figura 11.- Cumplimiento del acuerdo por euskaltegis en la PE1

En la Figura 12 se presentan los datos de los cuestionarios entregados y recogidos en cada uno de los euskaltegis. En (Arruabarrena y Pérez, 2005a) se hallan tablas más exhaustivas sobre estos datos.

De los 22 centros que acordaron participar, 2 de ellos abandonaron (los centros 10 y 16, Figura 12), y 6 de ellos entregaron menos cuestionarios que los comprometidos (los identificados con los números 2, 6, 8, 12, 17 y 22).

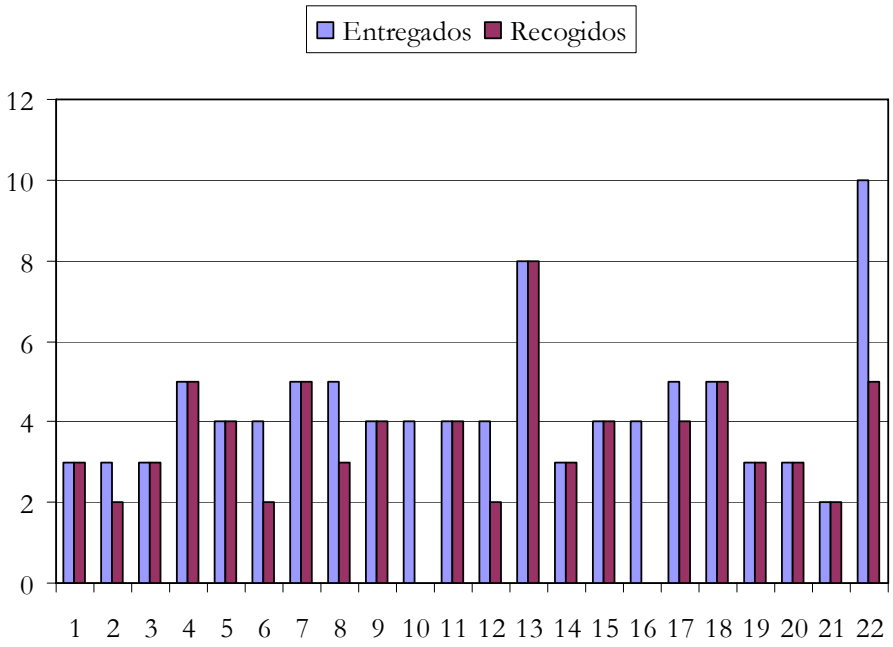


Figura 12.- Cuestionarios por euskaltegis de la PE1

En la Figura 13 se muestra la evolución de la recogida de los cuestionarios. Se puede observar que en el período que iba desde junio a septiembre no se obtuvo ningún cuestionario. Se recogieron el 78% de los cuestionarios entregados. Aunque el 91% (27+64) de los centros entregaron cuestionarios completados, estos centros entregaron de media 0.6 cuestionarios menos de los que se habían comprometido.

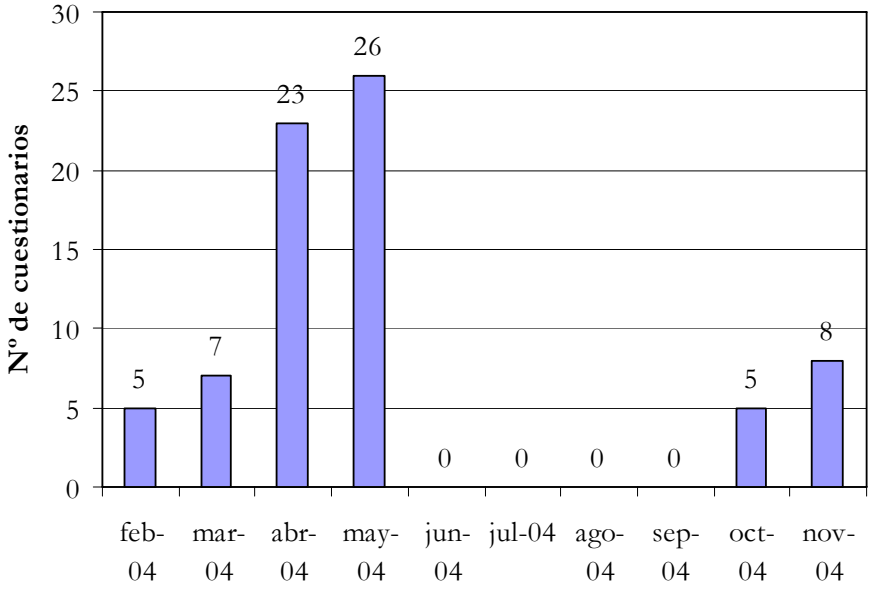


Figura 13.- Número de ejemplares de cuestionarios recogidos en la PE1

Con respecto a los ítems (Figura 14) se recogieron 9 valoraciones en 37 ítems (15%), 10 valoraciones en 179 ítems (71%), 11 valoraciones en 24 (10%) y 61 valoraciones en los 12 ítems de

anclaje (5%). Esto suponía que un **85% del banco de ítems tenía el número de valoraciones** que se había **fijado** obtener (10), y el resto una valoración menos. En todos los casos se superó las 7 valoraciones recomendas.

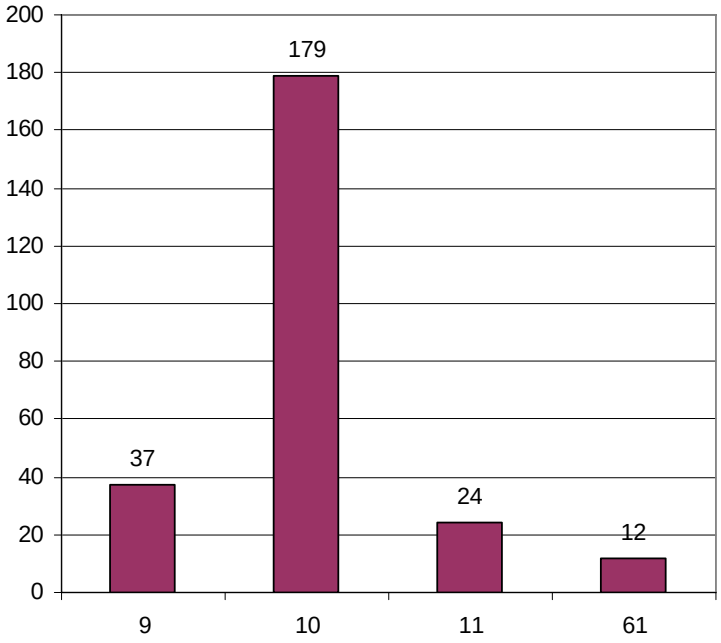


Figura 14.- Valoraciones por ítem recogidas en la PE1

6.1.6. Incidencias

Cinco meses después de comenzar la PE1, se observó que tan solo se habían recuperado 61 cuestionarios completados frente a los 80 que se pretendían haber recopilado. La distribución de los ejemplares recuperados por número de cuestionario puede verse en la Figura 15. En ella se aprecia que del cuestionario 7 tan solo se habían recuperado cinco ejemplares, siete de los cuestionarios 5 y 6, ocho de los cuestionarios 3, 4 y 8 y nueve ejemplares de los cuestionarios 1 y 2.

Para paliar el déficit en el número de los cuestionarios recopilados, y con vistas a concluir con la prueba de campo, se optó por componer varios cuestionarios ad hoc, de la misma longitud que los iniciales, pero con los ítems menos respondidos hasta aquel momento. Como consecuencia se recuperaron 13 de los 14 cuestionarios creados (Figura 13), dando por finalizada con ello la prueba de campo.

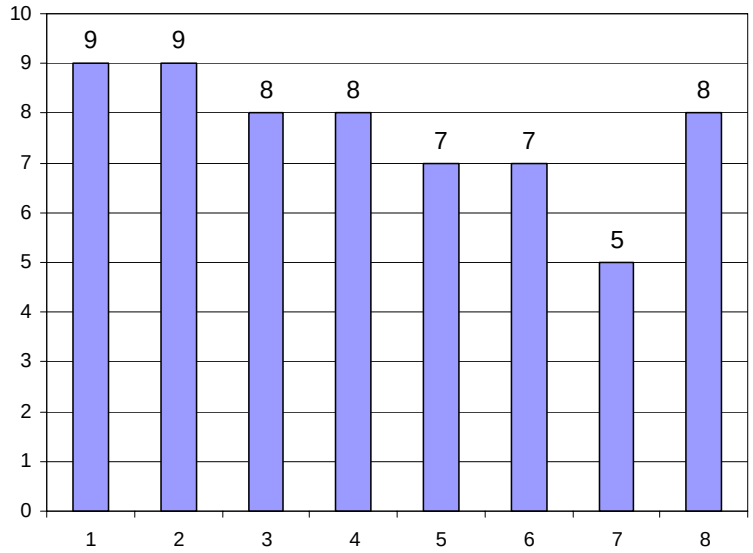


Figura 15.- Distribución de los 61 primeros ejemplares de cuestionarios recuperados por número de cuestionario

Además, como consecuencia de haber empleado cuestionarios en formato papel, hubo expertos que no respetaron las instrucciones de completado de los mismos como se ve en la Figura 16. De las 3119 aportaciones sobre los ítems que se recopilaron, un 84% de ellas (2640) siguieron las instrucciones indicadas. Sin embargo, el 2% de las valoraciones (53) omitieron alguno de los valores solicitados, y el 1% (20) las dos. Por el contrario, otro 1% (18) aportó más de una respuesta a uno de los rasgos, y el 12% restante (388) aportaron múltiples respuestas a ambos rasgos consultados. Posteriormente, durante la fase de análisis de datos y calibración de ítems, hubo que tomar medidas correctoras al respecto.

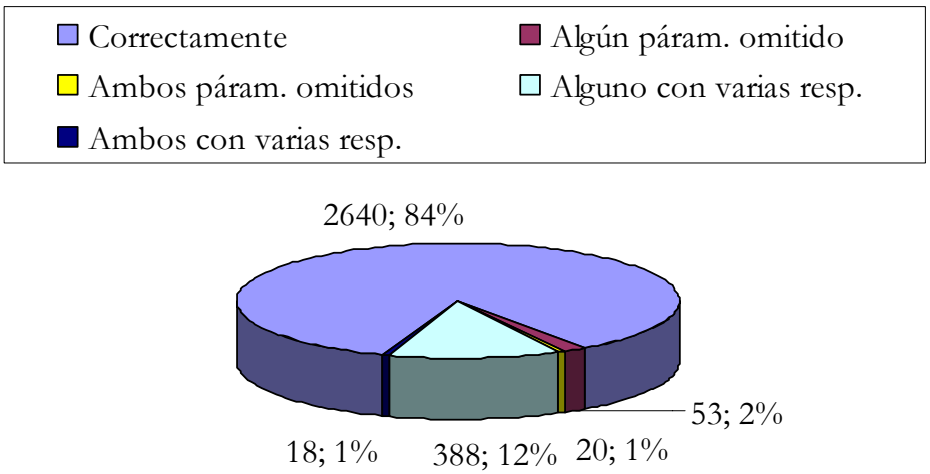


Figura 16.- Cumplimiento de las instrucciones de completado por parte de los expertos de la PE1 en los parámetros de los ítems

6.1.7. Mejoras

En este apartado se enuncian las mejoras que podrían acometerse en futuras réplicas de la PE1. Las mejoras identificadas son fruto del seguimiento, la medición y el análisis del desarrollo de la prueba.

Un aspecto a tener muy en cuenta es que el desarrollo de la prueba de campo *no* tenía que *interferir con la actividad profesional de los sujetos pasivos*. Este hecho era más importante aún cuando su colaboración era voluntaria y no remunerada. Para futuras instanciaciones, se propone **sincronizar el comienzo de la prueba de campo con todos los participantes**. De esta manera, no se alargará el desarrollo de la prueba. Del mismo modo, a la vista de los costes que ha supuesto, se plantea también **reducir el número de desplazamientos hasta los centros** para entregar y recoger cuestionarios. Esto ahorrará tiempo a los participantes activos y reduciría los costes.

Sin embargo, estas propuestas pueden redundar en una tasa de abandono mayor, de manera que, si se atienden, será conveniente preverla, más aún sabiendo que en esta prueba el abandono ha superado el 20% previsto. Finalmente, se considera conveniente **reducir el número de cuestionarios a comprometer** por centro ya que ello facilitará la aceptación del acuerdo y redundará en plazos más breves.

Para finalizar, puesto que la escala de medida era la misma en todos los cuestionarios (los niveles de HABE) y que en los cuestionarios se incluía una tabla de equiparación de los niveles de dificultad diferentes perfiles lingüísticos y certificados oficiales, se vio que no era necesario realizar equiparación de las valoraciones de los expertos. Por ello, y si las condiciones se repiten, se recomienda **eliminar los ítems de anclaje de todos los cuestionarios** y hacer un reparto de estos ítems como se ha procedido con los demás. Esto reducirá el número de cuestionarios.

También se propone **modificar las instrucciones de completado del cuestionario** para resaltar que **las valoraciones dadas por los expertos deben limitarse a escoger una única opción entre las múltiples presentadas**.

Por algunas explicaciones dadas por los expertos que escogieron más de un nivel de dificultad, se sugiere **incluir en las instrucciones que se requiere que se indique una valoración neutra del nivel de dificultad** (frente a una valoración pesimista u optimista).

Por la experiencia realizada, cuando no hay duda sobre el rango de valores entre los cuales el experto debe decantarse a la hora de emitir su valoración, **el cuestionario no debe dar opción al experto de responder un valor no incluido en el rango de valores considerados**. Trasladando al caso de Hezinet, la opción "otra destreza" no debiera haber estado incluida en los cuestionarios entre las opciones a barajar por los expertos administrados.

6.1.8. Evaluación de costes

En la Figura 17 se muestra el **tiempo invertido** en el desarrollo de la prueba por los participantes activos y pasivos. Se invirtieron 128h en formación para realizar la PE1. Además, las tareas de planificación y gestión supusieron 108h. La conducción de la prueba fue la etapa más extensa ocupando 302h de los participantes activos y 77h de los pasivos. En términos medios, un experto necesitó 50 minutos para elaborar su aportación. Finalmente, la elaboración de informes consumió 164h. Todos estos apartados sumados ascendieron a 779h.

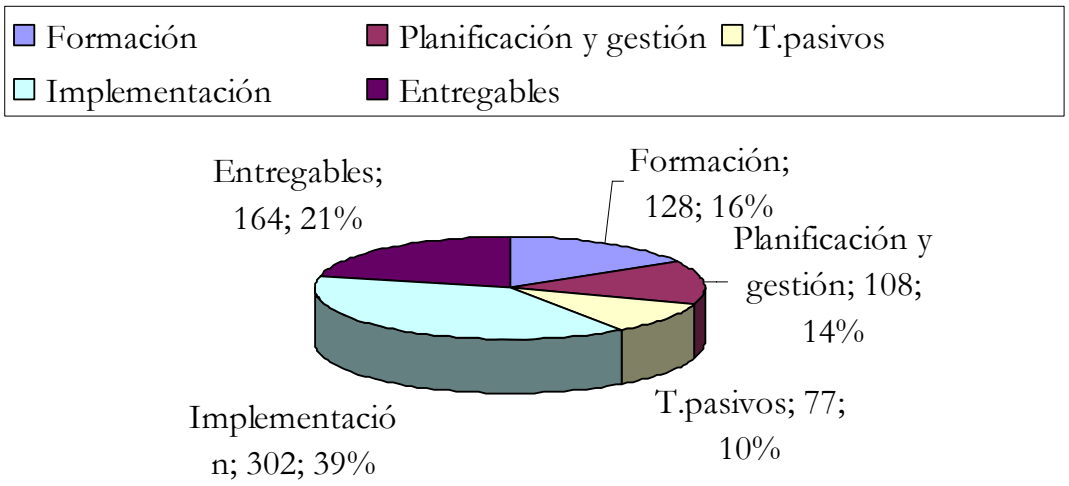


Figura 17.- Tiempo invertido en el desarrollo de la PE1

En cuanto a **llamadas telefónicas** (Tabla 8), durante el desarrollo de la PE1 se realizaron 83 con una duración total de 362 minutos

(6h y 2min). Se enviaron 32 correos electrónicos, uno por centro consultado y al inicio del proceso de captación. En términos medios, se hicieron llamadas telefónicas con una duración total de 16.5min repartidas en 3.8 llamadas y se enviaron 1.5 emails por centro acordado. Desde el punto de vista de cuestionarios, el coste de recuperar un cuestionario supuso realizar 1.1 llamadas, estar en contacto directo durante 4.9 minutos y enviar 0.4 correos electrónicos.

LLAMADAS TELEFÓNICAS & EMAILS	**En total**	**Centro acordado**	**Por cuest.**
N. llamadas a centros a consultados	83	3.8	1.1
T. en llamadas con centros consultados (min)	362	16.5	4.9
N. de emails	32	1.5	0.4

Tabla 8.- Consumo en llamadas y correos electrónicos en la PE1

Desde el punto de vista de costes por **desplazamiento** en la PE1 (Tabla 9), se efectuaron 53 viajes a centros acordados, recorriendo un total de 1576km e invirtiendo en ello 2260 minutos (37h 40min). En términos medios, y desde el punto de vista de centros acordados, se realizaron 71.6km por centro invirtiendo en dicho desplazamiento 102.7 minutos (1h 42min) en 2.4 visitas. Obsérvese que, aunque sólo se acudía a los centros a recoger los cuestionarios tras haberse asegurado telefónicamente que los mismos estaban completados, el número medio de visitas fue mayor que 2, porque no en todos la recogida de los cuestionarios resultó exitosa y hubo que repetir la visita. Desde el punto de vista de cuestionarios, el coste de recuperar un cuestionario supuso realizar 0.7 viajes, 30.5 minutos de desplazamientos y 21.3km de desplazamiento. En (Arruabarrena, López-Cuadrado et al., 2007) se puede encontrar con mayor detalle información sobre los costes del desarrollo de la prueba.

DESPLAZAMIENTOS	**En total**	**Centro acordado**	**Por cuest.**
N. viajes a centros acordados	53	2.4	0.7
T. desplazamientos a c. acordados (min)	2260	102.7	30.5
Km.s desplazados a c. acordados	1576	71.6	21.3

Tabla 9.- Consumo en desplazamientos de la PE1

Para evitar interferir con los periodos de gran dedicación de los expertos, la PE1 comenzó a principios de año y se previó que finalizase a comienzos del verano. Sin embargo, la escasa

disponibilidad financiera, la necesidad de personal euskaldun para las interlocuciones con los euskaltegis y de desplazarse a los distintos centros para entregar y recoger en mano los cuestionarios limitó el número de participantes activos que podían desarrollar la labor a una única persona. Por este motivo, los acuerdos con los distintos coordinadores de centros, y sus posteriores entregas y recogidas se fueron realizando de manera secuencial, alcanzando el periodo estival, y extendiéndose hasta noviembre. La duración total de la prueba fue de 10 meses, aunque la duración real se reducía a 6 si se eliminaban los 4 meses mencionados anteriormente.

En (Arruabarrena, López-Cuadrado et al., 2007) se recogen con más detalle los costes temporales y los asociados a llamadas telefónicas, correos electrónicos y desplazamientos. A estos costes, habría que añadirles los costes de las copias de los cuestionarios y un ordenador con software de ofimática.

6.2. Pruebas con expertos 2 (PE2)

El objetivo de esta segunda prueba se reajustó a obtener **7** nuevas **valoraciones por cada ítem**. El ajuste se hizo teniendo en cuenta la duración de la PE1, los recursos económicos y humanos disponibles y considerando las recomendaciones de los trabajos de (Dalkey, Brown et al. 1970) y (Shneiderman, 1998) (véase el apartado 5.3). Con los datos de ambas pruebas de campo, en promedio, se superarían las 15 valoraciones por cada ítem, lo que favorecería la significancia estadística de los resultados.

6.2.1. Participantes

Los sujetos participantes en la prueba fueron los mismos que los de la prueba PE1, con excepción de **los revisores**. **Se descartaron** estos sujetos, puesto que los ítems eran los mismos y no se había modificado la estructura de los cuestionarios, con lo que **no era necesario realizar pruebas piloto** al seguir siendo válidas las realizadas en la PE1.

Se estableció involucrar a 40 euskaltegis que acordaran completar 2 cuestionarios cada uno de ellos, lo que implicaría a 80 **expertos** evaluadores.

6.2.2. Metodología

Como esta prueba de campo era similar a la primera y los objetivos eran los mismos, el proceso de recogida de datos fue similar. Tan solo se modificó el procedimiento para incluir algunas de las propuestas de mejora obtenidas de la prueba anterior. En los siguientes puntos se incluyen las principales diferencias:

- El diseño se heredaba, pero **se modificaba la distribución de los 252 ítems en los cuestionarios**. En concreto los cuestionarios se reducían a 6 con 42 ítems cada uno, no habiendo ítems de anclaje.
- Las pruebas piloto de la PE1 eran válidas en esta prueba de campo, en consecuencia **no era necesario realizar nuevas pruebas piloto**, **ni se precisaban** participantes pasivos con el rol de **revisores**, dado que se administrarían los mismos ítems y la estructura de los cuestionarios sería la misma que en la prueba anterior.
- Se mantendría una primera llamada de captación a los participantes expertos para acordar su participación y que estuvieran sobre aviso del **envío de los cuestionarios**.
- Se previó una **tasa de abandono** de los expertos del 50%. Se estimó que la no entrega en mano de los cuestionarios redundaría en un menor compromiso por parte de los participantes voluntarios.
- El número de cuestionarios por centro se reducía a dos.
- La **administración** se haría del mismo modo que en la PE1 salvo que la **entrega y recogida de cuestionarios se haría por correo postal.** Esta opción permitía que los distintos centros pudieran realizar la tarea en paralelo (vs escalonadamente como en la PE1), sin tener los problemas que suponían la entrega en mano. Además, por el mismo importe postal, se podía extender la administración de la prueba a otras provincias limítrofes.

Al igual que en la PE1, se hizo **seguimiento** exhaustivo de los **costes invertidos** y **aspectos mejorables** durante el ciclo de vida de la PE2.

6.2.3. Resultados

Esta prueba se desarrolló entre octubre de 2004 y enero de 2005, recogiéndose un total de 42 cuadernillos rellenados provenientes de 20 centros. Durante el desarrollo de la PE2 (Tabla 10) se solicitó a 51 euskaltegis la participación en el estudio. De estos centros accedieron 40, acordando completar 81 cuestionarios.

MUESTRA	Consultados	Acordados	Reales
N. expertos		81	42 (52%)
N. de centros (euskaltegis)	51	40	20 (50%)
Media de expertos x centro		2.0	2.1 .

Tabla 10.- Tamaño de la PE2

Sin embargo, durante el desarrollo de la prueba, el 48% (100-52%) de los expertos abandonó, por lo que finalmente se recogieron 42 cuestionarios de 20 centros (el 50% de los centros acordados). En términos porcentuales, tal y como refleja el gráfico de la Figura 18, un 50% de los centros cumplió el nivel de sus compromisos (uno incluso lo incrementaba) y un 50% abandonó el estudio.

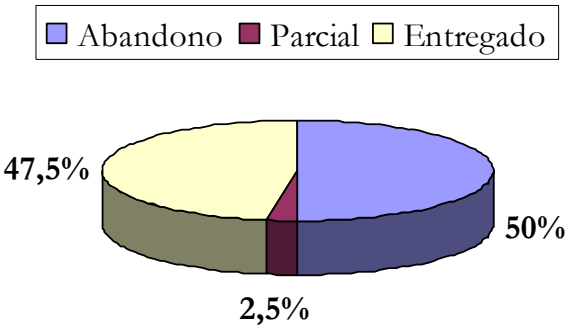


Figura 18.- Cumplimiento del acuerdo por euskaltegis en la PE2

En términos medios, se entregaron 2 cuestionarios por centro y se recogieron 2.1. El que se hubieran recogido, en término medio, más cuestionarios que los entregados se debió a que dos centros (11 y 16 en la Figura 19) entregaron tres cuestionarios, el primero porque se había comprometido a ello y el segundo porque completó otra copia de uno de los dos cuestionarios que había recibido. En la Figura 19 se presentan los datos de los cuestionarios enviados y recogidos en

los euskaltegis. En (Arruabarrena y Pérez, 2005b) se hallan tablas más exhaustivas sobre estos datos.

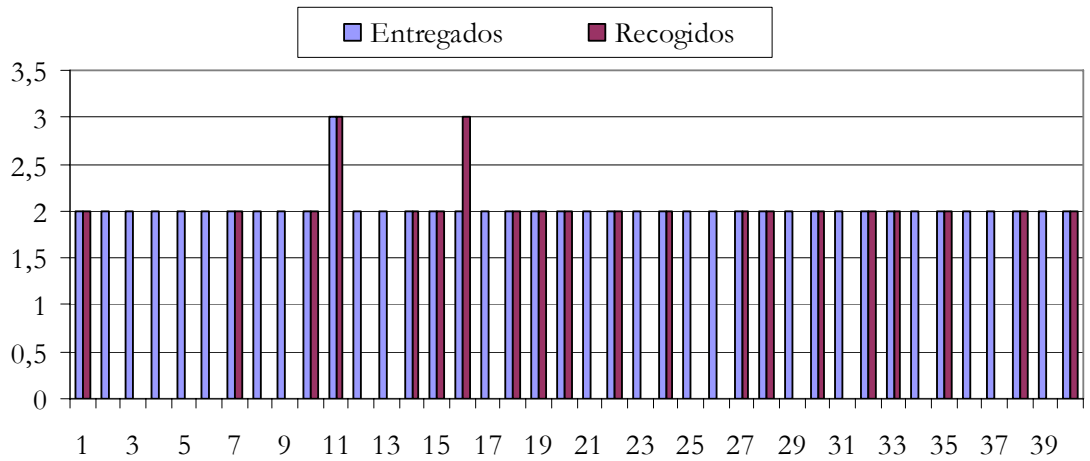


Figura 19.- Cuestionarios por euskaltegis de la PE2

De los 40 centros que acordaron participar, 20 de ellos (50%) abandonaron (los centros 2-6, 8-9, 12-13,17, 21, 23, 25-26, 29, 31, 34, 36-37 y el 38). Uno de ellos (2.5%) entregó más cuestionarios que los comprometidos (el 16). En total se recogieron el 52% de los cuestionarios entregados. Esta cifra fue algo mejor que la prevista (50% de los entregados).

La distribución de los cuadernillos recuperados por tipo de cuestionario se halla en la Figura 20. Si bien el objetivo de recuperar 7 ejemplares se había igualado o superado con los cuadernillos tipo 1, 2 y 5 (con 7, 9 y 8 ejemplares recuperados respectivamente), no había sucedido lo mismo con los cuestionarios números 3, 4 y 6, de los cuales solo se habían recuperado 6 ejemplares. Sin embargo, no se tomó ninguna medida correctora, procediendo con el análisis de los datos.

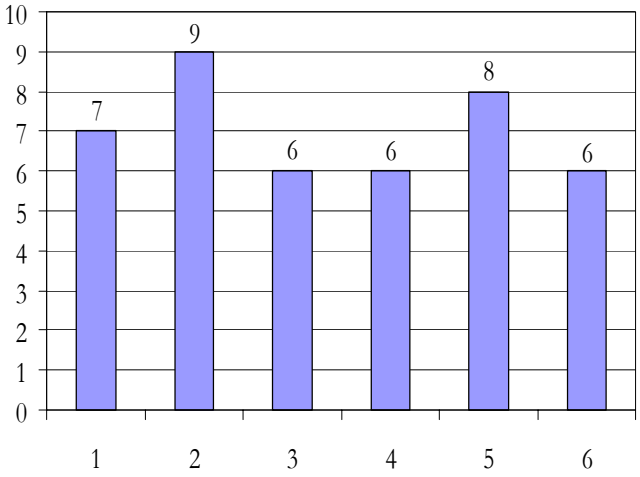


Figura 20.- Distribución de los 42 ejemplares de cuestionarios recuperados por número de cuestionario

En la Figura 21 se muestra la evolución de la recogida de los cuestionarios durante la ejecución de la PE2.

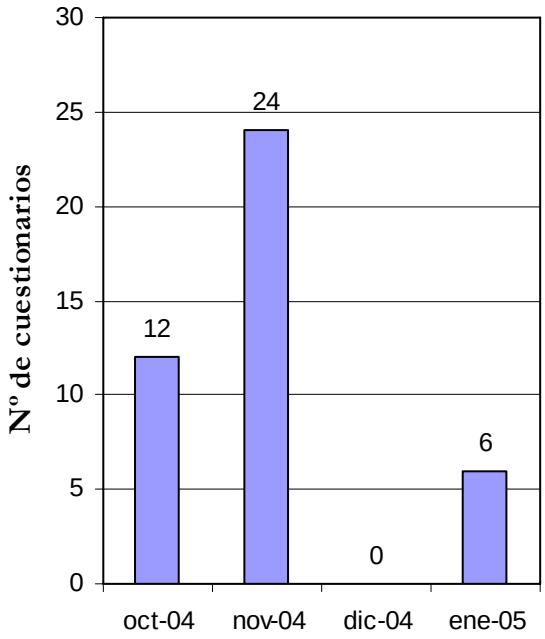


Figura 21.- Número de ejemplares de cuestionarios recogidos en la PE2

Se observa como el mes de diciembre no fue productivo en cuanto a cuestionarios recibidos. También se puede observar el patrón de comportamiento en los distintos centros: desde los que respondieron casi inmediatamente a la petición (véase cuestionarios recogidos en octubre) a los que se tomaron el tiempo previsto para responder (noviembre) y los que respondieron tras las vacaciones de navidad (enero).

Con respecto a ítems (Figura 22), se recogieron 4, 10 y 11 valoraciones respectivamente, para tres ítems aislados (1%). Hubo 3

ítems con 5 valoraciones (1%), 6 ítems con 13 (45%), 7 con 53 (21%), 8 con 41 (16%) y 9 con 39 (15%). Esto supuso que el 99.6% del banco de ítems tuviera 5 o más valoraciones, el número mínimo recomendado por (Shneiderman, 1998), y el 54% tuvo al menos las 7 valoraciones que se quería obtener mediante la PE2 y que (Dalkey, Brown et al., 1970) consideran como número óptimo de contribuciones de expertos. **El promedio de las valoraciones conjuntas** de ambas pruebas de campo por ítem eran iguales o superiores a 15, habiéndose **alcanzado** así el **objetivo** de esta prueba.

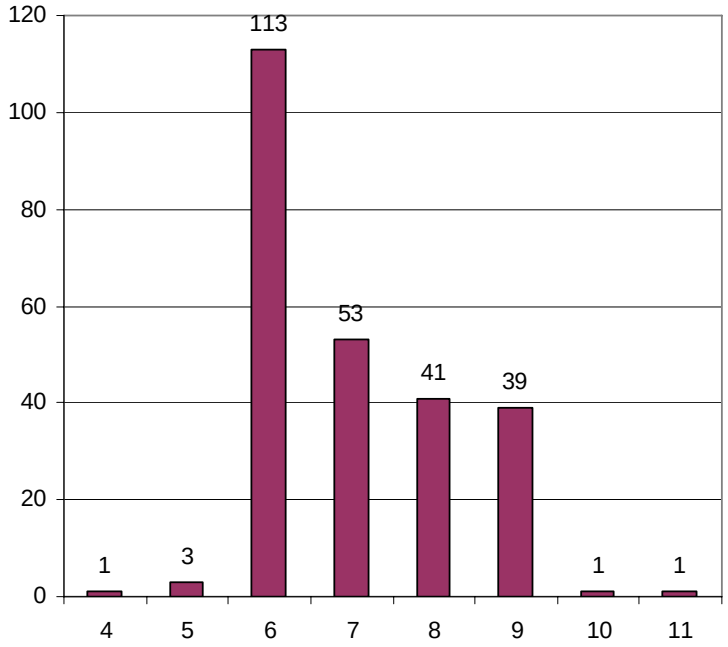


Figura 22.- Valoraciones por ítem recogidas en la PE2

A la vista de los resultados, y dando un paso más en la interpretación de los datos brutos recopilados, hay que aclarar que el ítem con identificador 184 fue el ítem que menos juicios recibió (4) (mayor número de respuestas omitidas) y, por el contrario, los ítems con identificadores 181 y 204 fueron los ítems con mayor número de valoraciones recopiladas (11 y 10 respectivamente).

6.2.4. Incidencias

Dos meses después de haber enviado los cuestionarios y puesto que para diciembre de 2004 únicamente se habían recuperado 36 (el 44% de los enviados), se contactó con 21 euskaltegis vía correo electrónico instándoles a remitir de vuelta los cuestionarios

completados a pesar de que el plazo de compleción inicial ya había expirado. De este modo, en enero se recuperaron 6 cuestionarios más.

Si bien con algunos números de cuestionario no se llegó al objetivo (cuestionarios 3, 4 y 6), no se tomó ninguna medida de correctora.

Del mismo modo que en la PE1, el uso de cuestionarios de papel permitió que algunos expertos no hubieran respetado las instrucciones de completado (Figura 23). De las 1768 valoraciones sobre los ítems que se tendrían que haber recopilado, un 84% de ellas (1480) siguieron las instrucciones indicadas. Sin embargo, el 6% de las valoraciones (101) omitieron alguno de los valores solicitados, y el 3% (54) los dos. Por el contrario, el número de expertos que aportó más de una respuesta a uno de los rasgos (2) fue inapreciable (0.11%), y el 7% restante (131) aportaron múltiples respuestas a ambos rasgos consultados. Al igual que en el otro caso, se tuvieron que tomar medidas correctoras al respecto durante la fase posterior de análisis de datos y calibración de ítems.

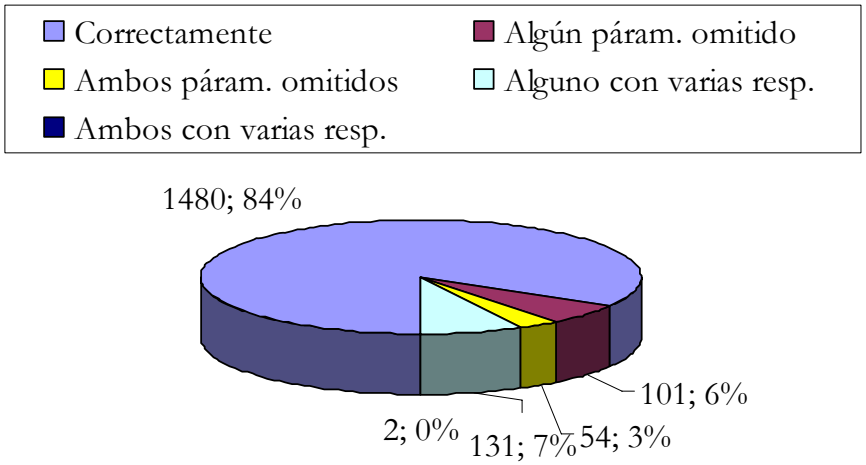


Figura 23.- Cumplimiento de las instrucciones de completado por parte de los expertos de la PE2 en los parámetros de los ítems

6.2.5. Mejoras

Al igual que en la PE1, y para futuras administraciones de la PE2, se sugiere **modificar las instrucciones de completado del cuestionario** para resaltar que **las valoraciones dadas por los expertos deben limitarse a escoger una única opción entre las múltiples presentadas.** La modificación podría incluir ejemplos

ilustrativos con respuestas no válidas junto con sus respectivas explicaciones del porqué de su rechazo. Asimismo, se propone **añadir en las instrucciones que se les requiere que indiquen una valoración neutra de nivel de dificultad**.

Se sugiere también que a la hora de cerrar el acuerdo de participación se proponga a los centros que completen 2 ó 3 cuestionarios y que estos decidan el número concreto que se les tenga que remitir, con el objetivo de que redunde positivamente en el número de cuestionarios recolectados por centro y en total.

6.2.6. Evaluación de costes

En la Figura 24 se muestra el **tiempo invertido** en el desarrollo de la prueba por los participantes tanto activos como pasivos. Se invirtieron 128h en formación para realizar la PE2. Además, las tareas de planificación y gestión supusieron 95h. La conducción de la prueba fue la etapa más extensa ocupando 248h de los participantes activos y 50h de los pasivos. En términos medios, un experto necesitó 50 minutos para elaborar su aportación. Finalmente, la elaboración de informes consumió 155h. La suma de los 5 apartados se elevó a 676h.

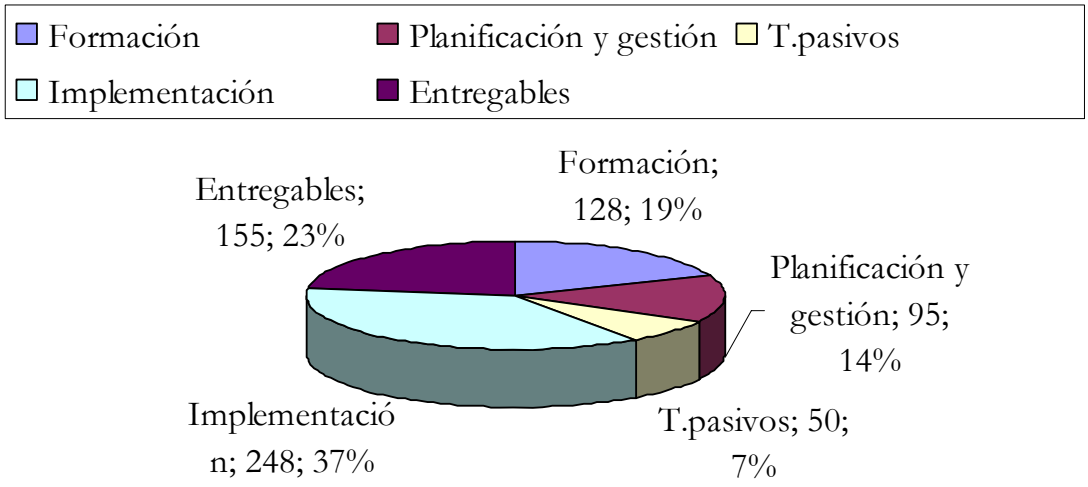


Figura 24.- Tiempo invertido en el desarrollo de la PE2

En cuanto a **llamadas telefónicas** (Tabla 11), durante el desarrollo de la PE2 se realizaron 94 llamadas telefónicas con una duración total de 476 minutos (7h y 36min) y se remitieron 51 correos electrónicos. En términos medios, se hicieron llamadas telefónicas con una duración total de 11.9min repartidas en 2.4

llamadas y se enviaron 1.3 emails por centro acordado. Y desde el punto de vista de cuestionarios, el coste de recuperar un cuestionario supuso realizar 2.2 llamadas, estar en contacto directo durante 11.3 minutos y enviar 1.2 correos electrónicos.

LLAMADAS TELEFÓNICAS & EMAILS	En total	Centro acordado	Por cuest.
N. llamadas a centros a consultados	94	2.4	2.2
T. en llamadas con centros consultados (min)	476	11.9	11.3
N. de emails	51	1.3	1.2

Tabla 11.- Consumo en llamadas y correos electrónicos en la PE2

En (Arruabarrena, López-Cuadrado et al., 2007) se recogen tablas más exhaustivas con los costes temporales y los asociados a llamadas telefónicas, correos electrónicos y desplazamientos. Al igual que en la PE1, a estos costes, habría que añadirles los costes de las copias de los cuestionarios remitidos (cada uno de 33 páginas) y un ordenador con software de oficina (similar a Word, Access y Excel).

Se cumplieron las previsiones fijadas. Por un lado, aumentó el número de abandonos por expertos, pero sin sobrepasar la tasa prevista (del 22% al 48%). Y desde el punto de vista del compromiso adquirido por los centros que no abandonaron, mientras que en la PE1 entregaron, en término medio, medio cuestionario menos que los acordados, en la PE2 entregaron 0.1 cuestionarios más. Por ello, en una posterior réplica tal vez se podría sugerir a los centros que completasen 2 ó 3 cuestionarios en lugar de proponer únicamente 2, y dejar que el centro opte si se ve o no en condiciones de completar 3 cuestionarios.

6.3. Análisis y calibración basada en juicios de expertos

En el momento de comenzar la segunda fase de la calibración basada en juicios de expertos, la información aportada por estos se hallaba transcrita y almacenada en una base de datos MS Access. El tamaño de dicha muestra inicial era de **4887 entradas** correspondientes a **116 expertos** participantes en las pruebas de campo PE1 y PE2, y realizada sobre un banco de **252 ítems**, **característica paramétrica de la muestra** que denotaremos

mediante el triple **(m=4887; e=116; n=252)**. La muestra puede consultarse en (Arruabarrena y Armendariz, 2008). En términos medios, contenía 17 respuestas de experto por ítem. Y por cada ítem se habían solicitado la respuesta al propio ítem, una valoración del nivel de dificultad y otra de la destreza lingüística trabajada por el mismo, ya que en base a estos dos aspectos tenía Hezinet organizados los ítems.

Este apartado se centra en la calibración de los ítems a partir de las aportaciones de los expertos obtenidas tal y como se han descrito anteriormente en este mismo capítulo. Se realizó una **calibración bietápica** de los ítems: primeramente se estimó el parámetro de *dificultad* y seguidamente el de la *destreza*, ya que es este primer parámetro con el que se va a contrastar la calibración basada en expertos con la calibración estadística.

Antes de proceder con la calibración propiamente dicha, se depuró la muestra. Se establecieron **tres niveles de filtrado** para descartar ítems inválidos (junto con todas sus aportaciones), para descartar expertos (junto con todas sus aportaciones) que no habían participado en el estudio con el rigor esperado y para descartar aportaciones anómalas o incorrectas. En consecuencia, la aplicación de los filtros podía disminuir los valores de algunos de los tres parámetros que caracterizaban la muestra (su tamaño (**m**), el número de ítems que contiene (**n**) y/o el número de expertos (**e**) cuyas valoraciones persisten).

El **estadístico M.dif** se aplicó ítem por ítem empleando los valores subjetivos de la variable otorgados por los expertos y que se hallaban en la muestra depurada. Esta aplicación dio lugar, por cada ítem, a la estimación del valor del **parámetro Dificultad** que caracterizaba al ítem y que correspondía a la variable dependiente del mismo nombre.

Para **garantizar el resultado de los procedimientos empleados**, en la etapa de diseño de los cuestionarios se optó por incluir en los cuestionarios una tabla de niveles de dificultad equivalentes y reconocidos entre distintos organismos de la Comunidad Autónoma Vasca. Con ello que se evitó llegados a este punto de desarrollo la necesidad de añadir un proceso de equiparación de puntuaciones una vez estimadas las dificultades.

Esta simple acción sirvió para *minimizar el posible sesgo entre los evaluadores*. Por otro lado, se realizó un **análisis de concordancia entre observadores** (**confiabilidad**) para determinar en qué grado los expertos estaban de acuerdo con la estimación calculada.

Finalizados los cálculos de la calibración con respecto al parámetro dificultad, y **para concretar las destrezas** a la muestra ya depurada para tal fin (tercer nivel de filtrado), se procedió de forma análoga pero aplicando el **estadístico M.est**. Así mismo, se realizó un **análisis de confiabilidad para determinar la fiabilidad de los resultados obtenidos**.

Los siguientes apartados identifican con mayor detalle los procesos ejecutados para realizar la calibración bietápica de ítems basada en aportaciones de expertos.

6.3.1. Depuración de la muestra

Puesto que había ítems marcados como "potencialmente erróneos" e incidencias registradas, se establecieron los siguientes criterios de filtrado para ítems y aportaciones de expertos en el orden en que se indican, siguiendo recomendaciones de depuración de marcos imperfectos (Frechtling y Sharp, 1997; Pérez, 1999):

Primera criba:

- *«C.ex2»*: Las aportaciones de expertos sobre los ítems se consideran válidas siempre que **indiquen solo un nivel de dificultad.**

Con este filtro se descartan las aportaciones que no tienen estimación de dificultad o tienen más de una, por lo que resultan inservibles para calcular la dificultad de los ítems.

Este criterio descartó el experto con identificador 234, que había optado por omitir sistemáticamente la dificultad de los 42 ítems que se le habían presentado, y descartó también otras 321 entradas más, de manera que la muestra se redujo a 4524 entradas (ver Tabla 12).

Segunda criba:

Compuesta por varios criterios para preparar la muestra para la posterior estimación del parámetro dificultad:

- *«C.it1»*: Por un lado, un ítem se acepta si **al menos un 50% de los expertos responde correctamente al mismo**.
- *«C.it2»*: Por otro lado, se mantienen aquellos ítems en los que **el 75% de las valoraciones de su nivel se encuentran agrupadas en 4 niveles consecutivos de dificultad**.
- *«C.ex1»*: Además, y en cuanto a expertos se refiere, se **eliminan los cuestionarios de expertos que no superan el 75% de respuestas a los ítems correctas**.

El objetivo del criterio C.it1, es eliminar ítems no fiables. Para ello se considera no fiables aquellos ítems con tasa de acierto inferior al 50%.

Por otro lado, el criterio C.it2, descarta aquellos ítems en los que no hay posibilidad de consenso ya que tienen una distribución de las valoraciones de dificultad muy dispersa.

Similarmente, el objetivo del criterio C.ex1 es eliminar aquellas contribuciones que no se hayan realizado con rigor por parte de los expertos. En nuestro caso, a priori se estableció una malla de cribado algo amplia, ya que se contaba con un número elevado de ítems "potencialmente peligrosos" (8.7% de los ítems; 22 de 252 ítems) que podían hacer que los mismos expertos tuvieran dificultades a la hora de responderlos. Ahora bien, una vez establecido que un experto no era fiable, se eliminaría de la muestra todas las aportaciones emitidas por éste, lo que descartaba su cuestionario.

Si bien estos tres filtros podían aplicarse una sola vez, se optó por aplicarlos reiteradamente hasta estabilizar los resultados, lo que sucedió tras dos aplicaciones sucesivas de los mismos. Para decidir el orden de aplicación de los tres criterios de filtrado se hicieron varias simulaciones. La opción seleccionada fue la que más valoraciones descartó.

La criba descartó 54 ítems, 4 expertos y 1209 valoraciones en total. Los identificadores de los ítems descartados pueden consultarse en el anexo A3 donde se hallan etiquetados con los descriptores de los

criterios que originaron su retirada del banco. La submuestra resultante del segundo cribado contenía información sobre 192 ítems, albergaba 3315 entradas (ver Tabla 12) y era la muestra que se emplearía para estimar la dificultad de los ítems.

Tercera criba:

Para depurar la muestra (m=3315; n=192; e=111) de valores imperfectos solo se le aplicó el siguiente filtro:

- *«C.ex3»*: se eliminarán aquellas valoraciones que **no tienen señalada una única destreza** en la respuesta de los cuestionarios.

Este criterio descarta aportaciones inservibles en la estimación del parámetro destreza de cara a hacer una *calibración de la destreza* de los ítems.

Se descartó añadir una nueva restricción, similar al C.it2, para eliminar ítems incorrectamente construidos basándose en la estimación de la destreza. La razón de ello fue que el uso de los ítems está orientado a la mejora de la capacidad lingüística de aquel sujeto que los trabaje y, a pesar de que la destreza estimada no sea una valoración mayoritaria entre las otorgadas por los expertos, ello no indica que el ítem no favorezca el desarrollo de dicha destreza.

La aplicación del criterio descartaba de la muestra entradas de expertos que hubieran omitido la respuesta, hubieran seleccionado múltiples destrezas entre las propuestas o bien hubieran escogido la opción "otra destreza". Concretamente, en la muestra inicial hubo un total de 369 aportaciones con un juicio de destreza distinto a las consideradas por Hezinet (7.6% de las aportaciones). El análisis de estos juicios mostró que estas valoraciones no se concentraban en algún grupo concreto de ítems sino que estaban dispersas entre los ítems del banco. Ambos apuntes, en opinión de la autora, reflejaban que el sistema de e-learning no necesitaba ampliar el rango de destrezas consideradas para integrar el nuevo banco de ítems calibrado.

Desde el punto de vista numérico, la aplicación de este último nivel de filtrado eliminó 439 valoraciones más, de manera que la submuestra resultante tenía 2876 entradas sobre 192 ítems emitidas

por 111 expertos (ver Tabla 12), siendo la submuestra que se emplearía para estimar las destreza de los ítems.

	M= n. de entradas	**N= n. de ítems**	**E= n. de expertos**
Inicial	4887	252	116
1er filtrado	4524 92.6%	252 97.2%	115 99.1%
2º filtrado	3315 67.8%	192 76.2%	111 95.7%
3º filtrado	2876 58.9%	192 .	111 .

Tabla 12.- Características de las submuestras de expertos consideradas durante la fase de depuración de los datos

Como **resumen de esta etapa** previa a la obtención de los parámetros, hay que subrayar que además de apartar aportaciones anómalas e incorrectas, se retiraron 60 ítems del banco (17.8%), y cinco cuestionarios completos de otros tantos expertos que optaron sistemáticamente por no responder a una de las cuestiones consultadas (el 4.3% de los expertos). Para realizar la calibración bietápica, los correspondientes análisis realizados sobre la muestra redujeron el volumen de ésta primeramente en un 32.2% (100-67.8; tercera fila en la Tabla 12) y posteriormente hasta un 41.1% con respecto al volumen inicial (100-58.9; última fila en la Tabla 12).

Consecuentemente, dado que este procedimiento podía suponer la eliminación de valoraciones, para garantizar que los estadísticos se aplicarían al número mínimo de aportaciones por ítems fijado, se concluyó que **la depuración de las aportaciones inservibles** (múltiples opciones de respuestas, omisiones,…) **debe realizarse antes de dar por finalizadas las pruebas de campo**. Las administraciones de las pruebas deben, además, considerar la necesidad de aumentar la recogida en 3 aportaciones más por ítem, en previsión de la merma del tamaño de la muestra por el descarte de los valores outliers a la hora de realizar las estimaciones.

Para mayor nivel de detalle sobre la aplicación de los filtros así como la repercusión de su aplicación en la muestra de aportaciones de experto puede consultarse (Arruabarrena y Armendariz, 2008). El informe recoge también un análisis sobre la incidencia de aplicar en órdenes alternativos las familias de criterios de filtrado aquí enunciados. Así mismo se argumentan más extensamente los criterios de depuración de la muestra junto con sus mallas de aplicación, los cuales debieran ser reconsiderados y ajustados, si

variasen los supuestos de fiabilidad de los ítems del banco a calibrar o bien el rigor con que participan los expertos.

6.3.2. Calibración de la dificultad

Para proceder con la estimación del rasgo dificultad, primero hubo que concretar el estadístico a emplear. Se desestimaron el *método Delphi* y el *algoritmo de máximo consenso* ya que ambas precisan juntar físicamente a varios expertos para que confronten sus opiniones.

En una primera aproximación se calculó la dificultad de los ítems aplicando directamente el estadístico habitual *moda*. Los resultados se pueden ver en el anexo A3 y en la Figura 25.

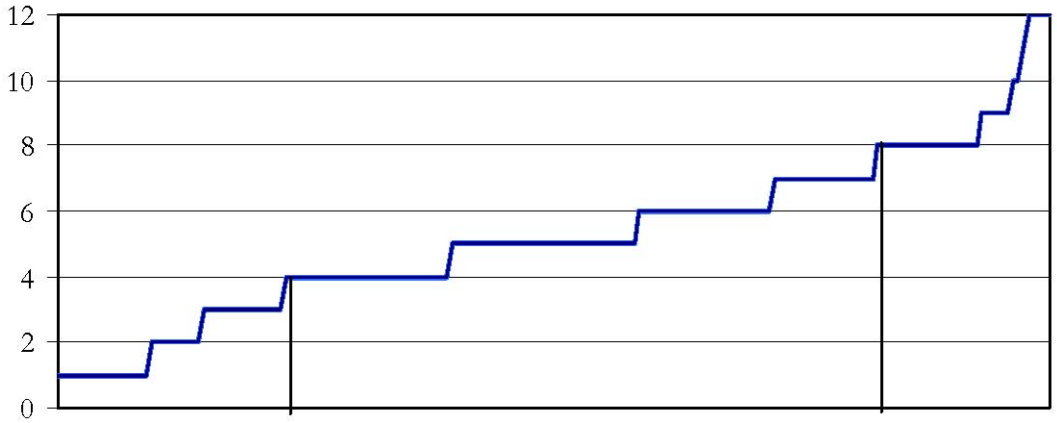


Figura 25.-Dificultad de ítems estimada por moda

En nuestro caso la moda no permitía determinar diferencias entre ítems del mismo nivel, y para ese objetivo se precisaba un estadístico que generase dificultades con valores no discretos.

Como segunda opción se aplicó la media para obtener valores continuos. Los resultados se pueden ver en el anexo A3 y en la Figura 26. Sin embargo, valoraciones outliers podían alterar la estimación, por lo que se desestimó emplear la media como estadístico para determinar la dificultad.

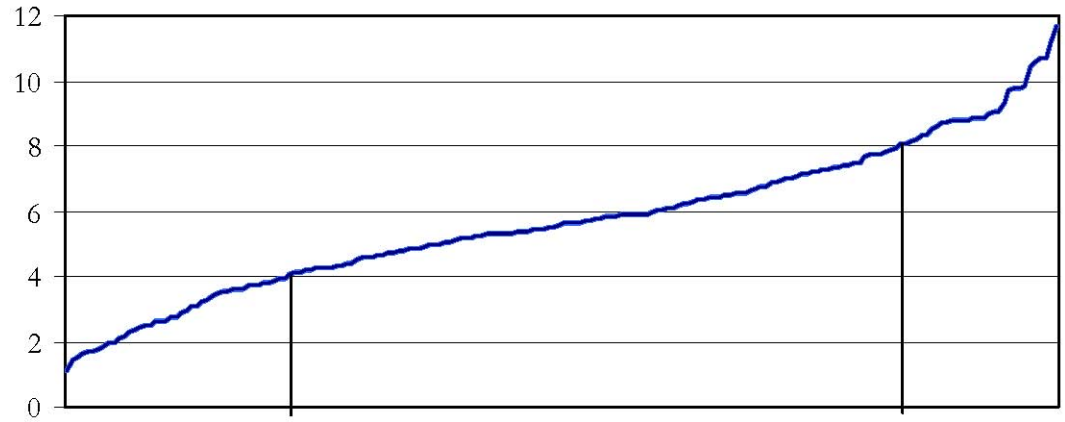


Figura 26.- Dificultad de ítems estimada por media

En su lugar se optó por buscar el consenso de los juicios de los expertos de forma off-line definiendo un procedimiento sistemático en el que el nivel de un ítem no se viera distorsionado por juicios extremos. Este procedimiento fusionaba la idea de máximo consenso y tenía similitud con la especificidad con que se había aplicado el criterio C.it2 de aceptación de ítems.

M.dif, el **estadístico dificultad del ítem**, se definió con dos reglas para establecer el valor más probable entre los juicios de dificultad más consensuados y genera un valor real en el intervalo [1-12]. **El objetivo del estadístico M.dif** era **establecer de forma off-line el valor más probable entre los pronósticos de dificultad más consensuados**, y constaba de las siguientes dos reglas:

- «M.dif1» la dificultad del ítem es el promedio de las frecuencias relativas de las valoraciones contenidas en el intervalo de 4 niveles (un tercio de la escala) con mayor densidad de valoraciones. Puesto que la muestra ya ha sido filtrada, al menos contendrá el 75% de las aportaciones.
- «M.dif2»: Si hubiera más de un intervalo que cumpla esa condición, entonces se extenderá el intervalo con un nivel más y se escogerá el intervalo con 5 niveles consecutivos con menor desviación.

A partir de la muestra resultante del segundo nivel de cribado y aplicando el estadístico *M.dif* se estimó la dificultad de los ítems. Los valores concretos estimados junto con sus desviaciones se han recogido en el anexo A3.

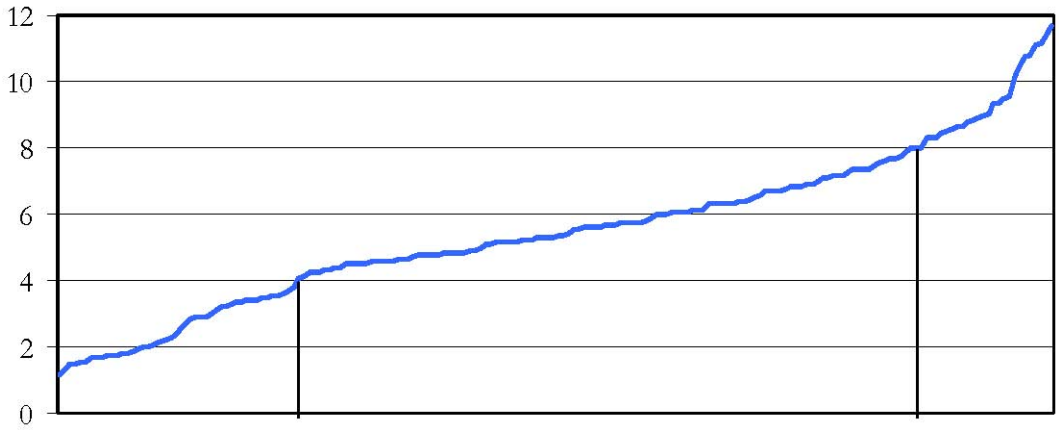


Figura 27.- Dificultad de ítems estimada por M.dif

Para facilitar al lector la comprensión del uso del estadístico M.dif, se han tomado como ejemplo los ítems 4 y 20. La Tabla 13 muestra las frecuencias relativas de las dificultades otorgadas por los expertos a estos dos ítems:

Ítem	Dificultad pronosticada												N. juicios
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	
4	8	3	1	2	1		1						16
20			1	6	4	3	1	1					16

Tabla 13.- Ejemplo de las valoraciones de dificultad otorgadas por los expertos a 2 ítems concretos

Según el estadístico M.dif1, para determinar la dificultad del ítem 4 se emplearían los 14 pronósticos del intervalo [1, 4] (que agrupaban al 87.5% del total de los pronósticos recogidos). La dificultad que se obtiene con dichos valores es 1.78, frente al valor 2.31 que generaría la media aritmética. Para el segundo ítem, puesto que existen dos intervalos consecutivos de 4 niveles que albergan ambos 14 aportaciones (el 87.5% de las emitidas), por aplicación de M.dif2, las valoraciones empleadas para calcular la dificultad del ítem son las del intervalo [3, 7] dando lugar a la estimación 4.8 y desviación 1.08, mientras que si se hubieran empleado las comprendidas en el intervalo [4, 8] el resultado que se hubiera obtenido sería 5.13 y desviación 1.24.

La regla M.dif2 se aplicó en 20 ocasiones (el 10.4% de las veces), ya que en ese mismo número de ocasiones se había producido la existencia de dos intervalos solapados de densidad máxima con cuatro niveles contiguos, por lo que el intervalo considerado para computar la dificultad se amplió a 5 niveles. En concreto, los identificadores de ítems comprendidos en esta casuística fueron: 20,

35, 49, 72, 76, 82, 87, 91, 99, 107, 159, 166, 167, 185, 192, 194, 198, 215, 222 y 231.

En la Figura 28 se muestran agrupados por niveles de dificultad discretos los 192 valores determinados por M.dif. La representación gráfica deja claro que aunque la escala de dificultad real tenía 12 niveles, prácticamente la mitad del banco de ítems tenía y tiene una dificultad estimada en alguno de los tres niveles 4, 5 ó 6. Concretamente el intervalo [4, 7) concentraba el 52.1% de las estimaciones de dificultad calculadas, y la ampliación al intervalo [1,9) abarcaba al 92.7% de los ítems calibrados. En cambio, apenas había ítems con estimaciones de dificultad elevada, de hecho el intervalo [11, 12] aglutinaba únicamente 4 ítems, siendo 11.6667 la estimación de dificultad más elevada del banco.

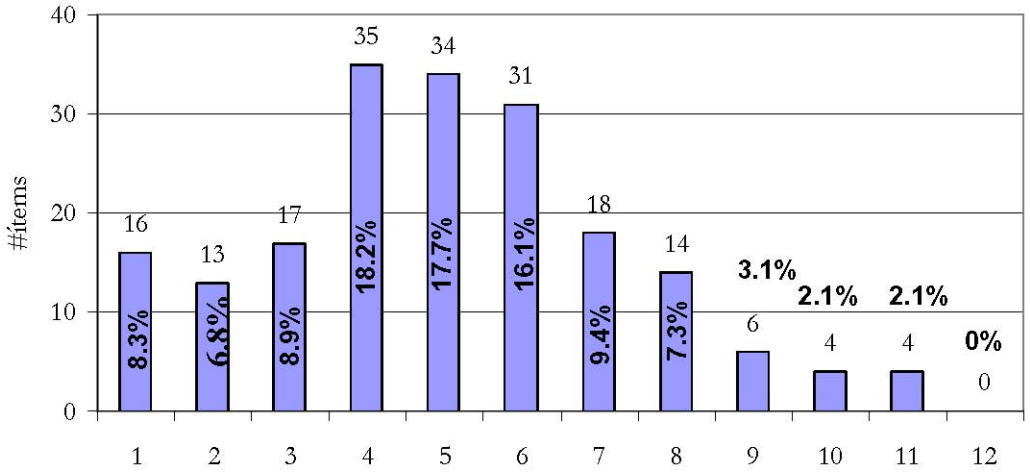


Figura 28.- Distribución de los ítems atendiendo a las dificultades estimadas (m=3315; n=192; e=111)

Así pues, y como **balance** de esta primera caracterización de los ítems, se puede indicar que la **distribución de las dificultades estimadas** de los ítems del banco fue **desigual**, de manera que la mitad del banco tenía una dificultad intermedia, y prácticamente el resto tenía dificultad media-baja, apenas habiendo ítem alguno con dificultad estimada elevada.

La estimación de las dificultades se realizó con un **índice de confiabilidad Kappa-Fleiss (κ) de 0.675**, ponderado con pesos cuadráticos, valor que (Fleiss, 1981) interpreta como ***buena*** **concordancia** entre los expertos.

En el anexo A3 se hallan especificados los 192 intervalos considerados por M.dif a la hora de computar la dificultad, los

valores de los índices de confiabilidad desglosados por ítem y contrastes entre las aplicaciones de los tres estadísticos utilizados para calcular la dificultad.

6.3.3. Calibración de la destreza

La segunda caracterización de los ítems se hizo empleando la muestra resultante del tercer filtrado. Para establecer la destreza de los ítems, primeramente se emplearon el *elemento mayoritario* y la *moda* como métodos de estimación.

La calibración mediante el elemento *mayoritario* aparece reflejada en la Figura 29. Esta opción no estimó la destreza de 10 ítems (un 6%): aquellos con frecuencias nominales inferiores al 51%.

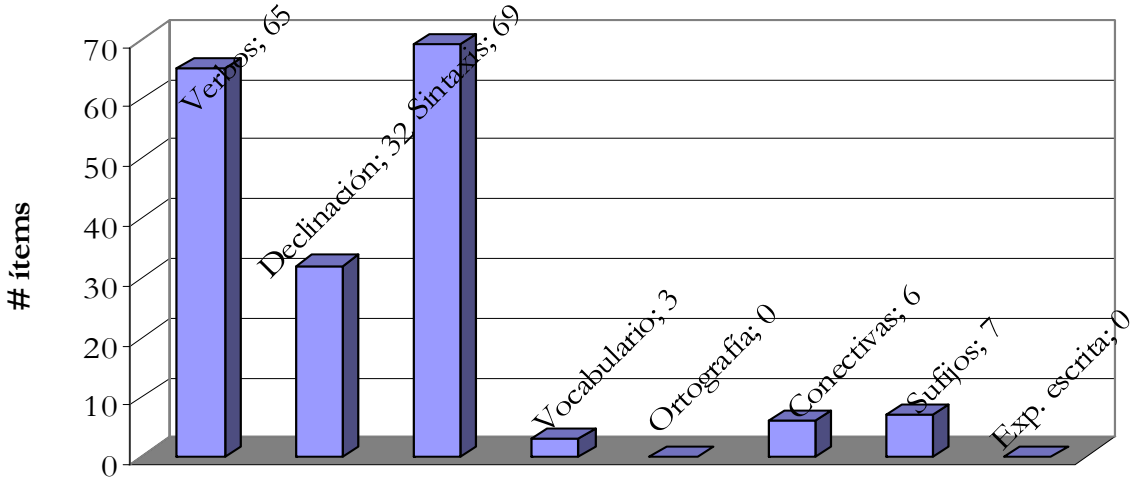


Figura 29.- Distribución de las destrezas estimadas por habilidad mayoritaria

La *moda* (Figura 30), aunque no marginaba ningún ítem, ofrecía 5 estimaciones múltiples (ítems con identificadores 11, 21, 98, 12 y 117).

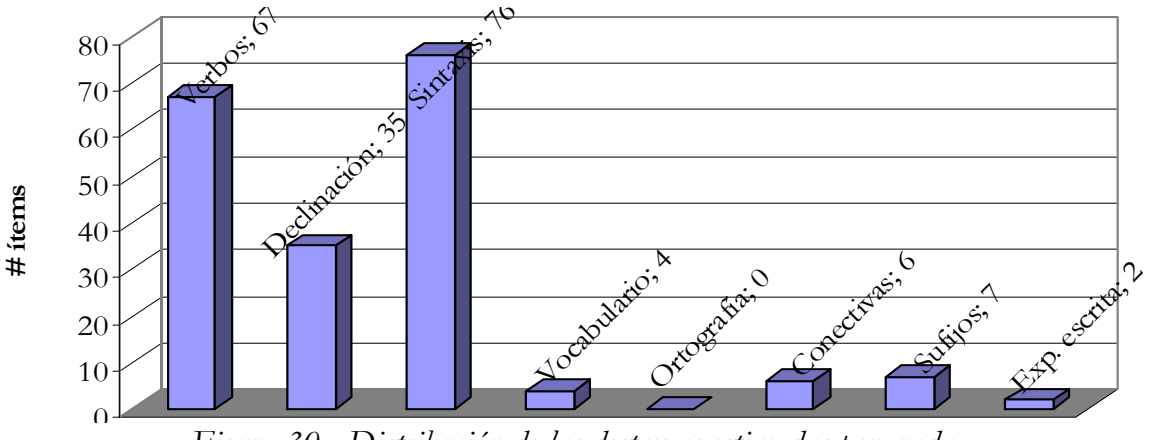


Figura 30.- Distribución de las destrezas estimadas por moda

Para nuestros objetivos era necesario concretar una única destreza para cada ítem. Por ello, se necesitaba desambiguar en caso de coincidencia de frecuencia entre las opciones más exitosas cuál era la destreza que caracterizaba al ítem. El **estadístico Destreza del Ítem** (**M.est**), definido por dos reglas, era una alternativa que resolvía esta situación. Su definición es la que sigue:

- «M.est1»: La destreza del ítem es la *moda* de las destrezas otorgadas.
- «M.est2»: Si hubiera más de una moda, se escogerá entre las mismas la menos frecuentemente trabajada por los ítems del banco, cuando ésta sea conocida.

La Figura 31 muestra la distribución resultante de las destrezas aplicando el estadístico *M.est*. Estas estimaciones junto con el volumen de expertos que coinciden con dichas atribuciones se pueden consultar en el anexo A3.

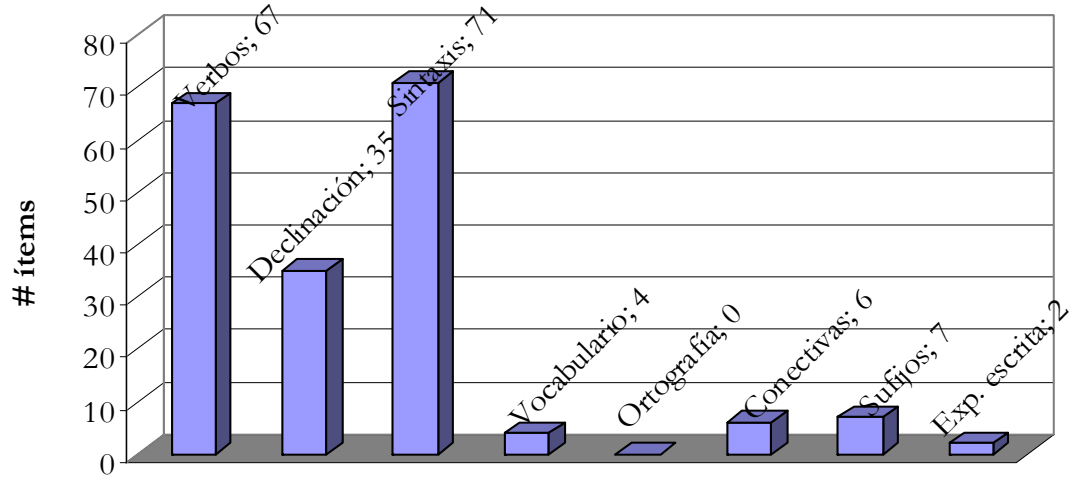


Figura 31.- Distribución de las destrezas estimadas por M.est

En concreto, un tercio de los ítems trabajaban la sintaxis, otro tercio los aspectos verbales, un sexto la declinación y el sexto restante el vocabulario, conectivas, sufijos o la expresión escrita.

La Figura 32 sintetiza el grado de acuerdo en las 192 estimaciones de las destrezas agrupándolas en secciones singulares. Concretamente, hubo *acuerdo pleno* de los expertos en el 35% de los ítems (67 ítems de los 192 del banco); para un 60% más de los ítems, la destreza estimada resultó ser por *acuerdo mayoritario*; y para el 5% por *acuerdo minoritario*.

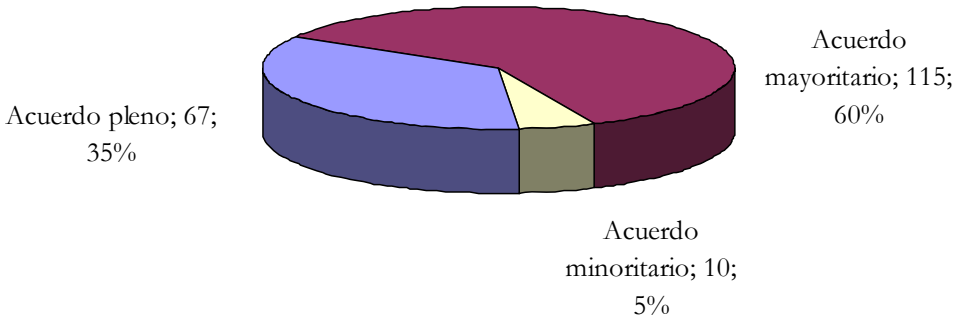


Figura 32.- Grados de acuerdo entre expertos en las estimaciones de destreza

La segunda regla del estadístico M.est se aplicó en los cinco casos de modas múltiples. En la Tabla 14 se recogen los casos, se indican el porcentaje de acuerdo entre los expertos, la destreza que se estimó según el estadístico y la destreza no contemplada. Concretamente en todos los casos se desestimó la opción de sintaxis a favor de otra destreza debido a que ésta era la destreza más frecuente.

N. ítem	% acuerdo	Destreza estimada	Destreza desestimada
11	50%	Declinación	Sintaxis
21	45%	Verbos	Sintaxis
98	30%	Vocabulario	Sintaxis
102	50%	Verbos	Sintaxis
117	50%	Declinación	Sintaxis

Tabla 14.- estimación de destrezas para modas múltiples

El **índice de confiabilidad Kappa-Fleiss (κ)** de la estimación de las destrezas de los ítems empleando los juicios de los expertos fue de **0.763**, valor que (Fleiss, 1981) considera **excelente**. Los valores de los índices de confiabilidad de cada uno de los ítems se pueden consultar en el anexo A3.

6.3.4. Evaluación de costes

Las **horas** utilizadas para realizar la calibración ascendieron a **287**. La Figura 33 recoge gráficamente el desglose de los tiempos invertidos: se emplearon 90 horas en formación, 55 en planificación y logística, 42 horas definiendo y aplicando los filtros y los estadísticos, 12 horas analizando resultados y 90 horas más elaborando documentación. En el apartado de planificación y gestión se incluyen las horas invertidas por dos asesores consultados.

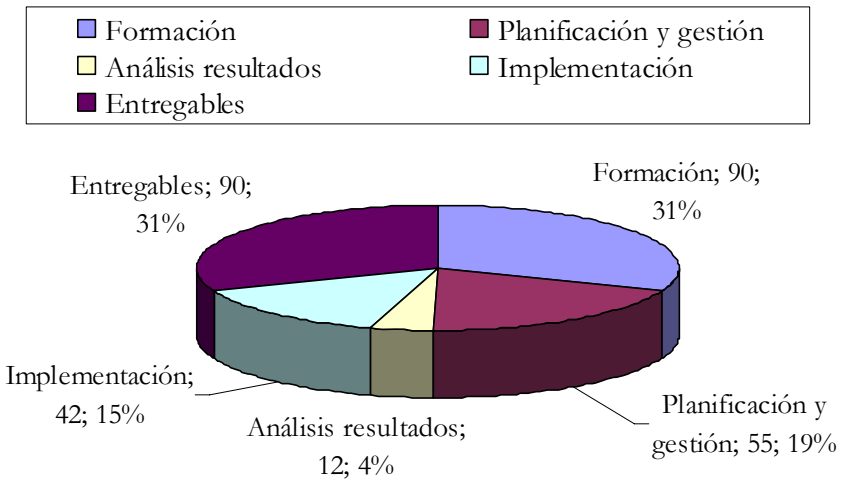


Figura 33.- Estimación del coste de horas invertidas en la fase "Análisis de datos y Calibración" de CE

Los **recursos utilizados** (véase Tabla 15), fueron un ordenador personal equipado con un paquete de ofimática (MS Office), el paquete estadístico SPSS y conexión a Internet. Además, hubo desplazamientos para tratar con los asesores.

Recursos económicos
Ordenador personal
Sw. de ofimática
SPSS
Conexión a Internet
Desplazamientos

Tabla 15.- Recursos empleados en la fase "Análisis de datos y Calibración" de CE

6.3.5. Análisis diferencial

La doble caracterización de los 192 ítems se realizó empleando las valoraciones conjuntas de todos los expertos. Desde el punto de vista de *validez de los resultados* los índices Kappa-Fleiss obtenidos corroboran el elevado nivel de acuerdo entre los expertos de ambas pruebas de campo.

Posteriormente, y también dentro de los estudios de fiabilidad, se analizó la repercusión de los criterios de filtrado en la depuración de las submuestras y se compararon las estimaciones tanto de niveles de dificultad como de destrezas por los expertos de la PE1 frente a los de la PE2, y todo ello para determinar en qué medida las dificultades/destrezas estimadas eran pronósticos consensuados por los expertos de la PE1, los de la PE2 o bien por todos los expertos. Estos estudios se pudieron realizar gracias a que las muestras

recabadas por cada prueba de campo contenían al menos 7 valoraciones por ítem.

En cuanto a la incidencia de los filtros sobre la muestra total y las submuestras, se compararon los volúmenes de las mismas en cuatros momentos específicos que fueron el volumen inicial tras terminar la recogida de datos (columna *inicial* Tabla 16 y Figura 34), y tras haber aplicado cada uno de los niveles de filtrado (últimas tres columnas Tabla 16 y Figura 34). Así, la recopilación de las aportaciones terminó con 3119 aportaciones de la PE1 y con 1768 de la PE2, de manera que en conjunto se habían obtenido 4887 aportaciones de las cuales el 63.82% correspondía a la primera prueba de campo y el 36.18% restante a la segunda (columna *inicial* en la Figura 34 y Tabla 16). Indicar que, si bien el número de aportaciones de las muestras fue decreciendo con cada nivel de filtrado aplicado, las proporciones de los volúmenes de aportaciones de los expertos de una prueba de campo y de la otra se mantuvieron prácticamente inalteradas. Específicamente, la **variabilidad** de las proporciones de los volúmenes de las submuestras sobre la muestra total **fue inferior al 1%** en los cuatros momentos considerados. Hay que puntualizar que durante la depuración de los datos también se descartó a 5 expertos, concretamente 2 de la PE1 y 3 de la PE2 que, como ya se ha comentado anteriormente, optaron sistemáticamente por no responder a alguno de los tres datos por ítem que se les solicitaba. A la vista de estos resultados, y sabiendo que se ejerció mayor control de seguimiento sobre los expertos de la PE1, cabe indicar que dicho esfuerzo no se ve reflejado en los datos brutos recogidos y depurados.

	Inicial	**1º filtro**	**2º filtro**	**3º filtro**
PE1	63.82%	64.26%	64.86%	64.64%
PE2	36.18%	35.74%	35.14%	35.36%

Tabla 16.- Volumen relativo de aportaciones de PE1 y PE2 frente al conjunto total durante la fase de depuración de los datos

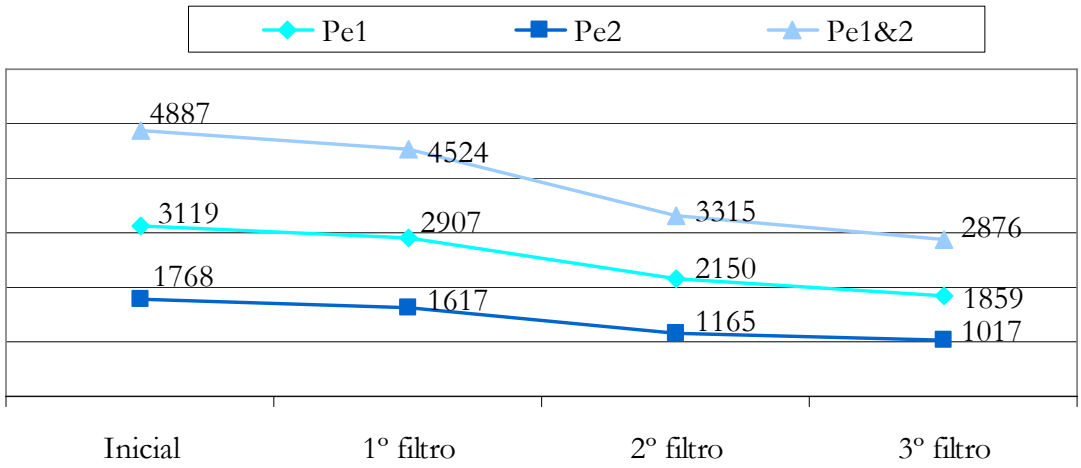


Figura 34.- Evolución del número de aportaciones de expertos por PE1, PE2 y ambas conjuntamente

Para **analizar el consenso en las dificultades estimadas**, se compararon los **intervalos de confianza** de la dificultad estimada por el estadístico M.dif con niveles de confianza del 95% y 99%, en lugar de comparar directamente los valores (continuos) de las dificultades estimadas. Para ello en primer lugar se procedió a calibrar paralelamente y de forma separada los 192 ítems aplicando M.dif, por un lado, sobre las 2150 aportaciones de la PE1 y, por otro lado, sobre las 1165 aportaciones de la PE2 (penúltima columna Figura 34). Seguidamente se computaron los IICC de M.dif empleando el coeficiente corrector de la t-Student para niveles del 95% y 99% de confianza. Finalmente, y por cada ítem, se analizó si había o no solapamiento entre los IICC calculados con los datos de la PE1 frente a los datos de la PE2. Concretamente, el **solapamiento** sucedía **en el 98% de los pares de intervalos cruzados,** alcanzando al **100%** de los pares de intervalos al aumentar **al 99% el nivel de confianza de los intervalos**. Considerando estos resultados y que el número de niveles de dificultad empleados era de 12, podemos concluir que **no hubo predominancia por parte de una submuestra de expertos en los valores de las estimaciones conjuntas**, por lo que estas últimas **podían considerarse estimaciones consensuadas de forma off-line**.

En la siguiente Figura 35 la línea azul continua refleja las dificultades estimadas empleando únicamente valoraciones la PE2, mientras que las líneas punteadas en el mismo color que se hallan paralelamente por encima y por debajo de la misma reflejan su

correspondientes IICC al 95%. Sobre la misma figura, en azul, se muestran también las estimaciones de dificultad para los mismos ítems empleando las valoraciones de la PE1 junto con sus respectivos IICC. El espacio entre tramas horizontales equivale a 2 niveles de dificultad desde el 0 al 12. Teniendo en cuenta todo esto, la figura se puede interpretar como si bien no hay coincidencia plena en las estimaciones de un grupo y otro de expertos, la semejanza es notoria. De hecho, en promedio, la diferencia dos a dos entre dificultades estimadas es de 0.41 (menor a medio nivel de dificultad).

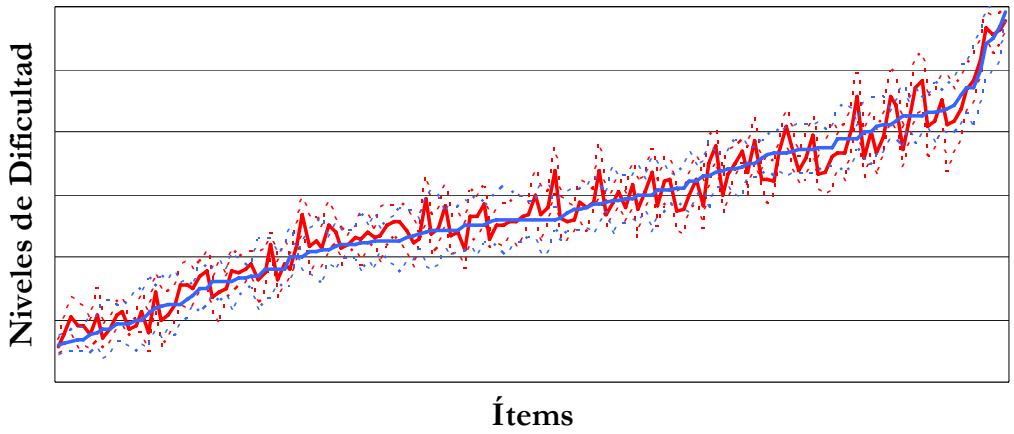


Figura 35.- Solapamiento de los IICC de M.dif aplicado a valoraciones de PE1 frente a valoraciones de PE2

Para realizar **la comparación de las destrezas estimadas** ítem a ítem, se compararon directamente las mismas **en términos porcentuales.** Se procedió de forma similar que para el rasgo dificultad. Así, a partir de las 1859 aportaciones de los expertos de la PE1 se aplicó M.est para concretar las destrezas según estos expertos, y se procedió de la misma forma con las 1017 valoraciones de los expertos de la PE2 (última columna Figura 34). Además, se calcularon las segundas opciones de destreza para aquellos ítems en que su primera opción de destreza no fue dada por una moda mayoritaria. El resumen de dichos cálculos se halla en la Tabla 17. Nótese que la coincidencia ítem a ítem de destrezas estimadas con juicios restringidos a una prueba de campo frente a la muestra total (PE1y2) fue del 94% (media de los valores 95.31% y 93.16% de la primera columna de la Tabla 17). Más aún, ésta se incrementó hasta alcanzar el 99% al comparar la 1ª y 2ª opción de las estimaciones de una prueba de campo con la primera destreza estimada de la muestra total. A la vista de estos resultados, se puede concluir que **no hubo predominancia** por parte de una submuestra **de expertos**

en los valores de las estimaciones conjuntas, por lo que estas últimas **podían considerarse estimaciones consensuadas de forma off-line**.

	Coinciden en Destreza		**Coinciden con 1ª o 2ª destreza**	
	#ítems	**vol.**	**#ítems**	**vol.**
PE1	183	95.31%	192	100%
PE2	177	93.16%	187	98.42%

Tabla 17.- Coincidencias ítem a ítem en destrezas estimadas PE1 vs PE1y2 y PE2 vs PE1y2

Los resultados de los tres últimos estudios de fiabilidad realizados indicaban que era **suficiente realizar una calibración empleando una muestra con 7 aportaciones por ítem** (una vez depurados los datos). También se pudo extraer como resultado que los estadísticos M.dif y M.est son capaces de proporcionar las estimaciones consensuadas de forma off-line, siempre y cuando exista un mínimo consenso entre los expertos participantes (índice de Kappa igual o superior a 0.5, por ejemplo).

6.4. Síntesis

En este capítulo se ha detallado el proceso llevado a cabo para calibrar los ítems de un banco a partir de las valoraciones otorgadas por expertos. La caracterización de los ítems se realizó atendiendo a los rasgos *dificultad* y *destreza* lingüística de los ítems, los mismos rasgos con los que organizaba y organiza sus ítems el sistema Hezinet, por lo que la calibración generada puede integrarse directamente en dicho sistema de aprendizaje para mejorar y ampliar su funcionalidad.

El coste de la calibración se ha calculado en términos de recursos humanos, económicos y temporales consumidos, empleando para ello las anotaciones registradas durante la realización de la calibración.

Para la recogida de datos se realizaron dos pruebas de campo, denominadas PE1 y PE2, que se ejecutaron de forma secuencial y haciendo uso de la técnica PDCA, de manera que la segunda ejecución fue una instancia mejorada de la PE1. Con la PE1 se recabaron, en promedio, 10 valoraciones por ítem y 7 más con la

segunda. En total participaron 116 expertos de 40 euskaltegis, que contribuyeron completando cuestionarios en formato de papel.

Para la depuración de los datos se definieron tres niveles de filtrado y para la calibración de los ítems los estadísticos M.dif y M.est con el objetivo de determinar de manera off-line el valor más probable entre los pronósticos más consensuados sobre los rasgos *dificultad* y *destreza* respectivamente. La depuración de la muestra que se realizó una vez recopilados los datos descartó 60 ítems y 5 expertos junto con sus contribuciones. Seguidamente, se realizó la calibración bietápica de los 192 ítems restantes del banco aplicando los estadísticos definidos.

Desde el punto de vista de *validez de los resultados* los índices Kappa-Fleiss obtenidos corroboran el elevado nivel de acuerdo entre los expertos de ambas pruebas de campo. Y los contrastes entre las contribuciones de los expertos de la PE1, los de la PE2 y la muestra conjunta en términos de intervalos de confianza y porcentuales corroboran que hubiera sido suficiente con realizar la calibración con una de las submuestras.

Así mismo, la experiencia del desarrollo ha servido para identificar de forma precisa los subprocesos embebidos y su organización temporal, así como las tomas de decisiones críticas que conforman una calibración con expertos. Estos conocimientos se emplearán para elaborar una propuesta de modelo de negocio de un sistema experto que ayudará en la toma de decisiones a la hora de desarrollar calibraciones de ítems. Algunas de las variables del modelo tendrán valores obtenidos de la calibración implementada.

A continuación se especifican las lecciones aprendidas de la puesta en marcha de las pruebas de campo con expertos y se remarcan algunos **aspectos** en los que hay que poner especial cuidado **para mitigar las amenazas** que pudieran malograr o invalidar los resultados en futuras instanciaciones:

- Si la colaboración de los expertos es voluntaria y no remunerada, hay que identificar su disponibilidad y adaptarse a sus compromisos laborales.
- Es importante conocer la "alfabetización informática" de los sujetos pasivos. La informática puede facilitar el proceso de

recogida de datos y su almacenamiento, pero también puede ser un inconveniente para que gente que desee participar no participe en el trabajo. Por otro lado los cuestionarios en papel pueden convertir una pregunta con opciones en una pregunta abierta (por ejemplo, añadiendo comentarios al margen), que se debe saber cómo tratar.

- La persistencia tanto en la captación de los expertos como en el seguimiento de cumplimentación de los cuestionarios, puede resultar contraproducente, pudiendo producir el efecto contrario.
- Las pruebas piloto son fundamentales. Aseguran que los sujetos comprenderán el objetivo, permiten conocer si el tiempo estimado para la cumplimentación es correcto y permiten ajustar con mayor atino la logística de la ejecución.
- Si la planificación temporal de la compleción de los cuestionarios ha sido validada en las pruebas piloto, hay que mantenerse inflexible a posibles cambios. Según nuestra experiencia, quien no cumple a tiempo lo acordado en una primera instancia, en la mayoría de los casos o bien termina por abandonar o bien su participación desde el punto de vista coste/beneficio no es rentable.
- Es más rentable, también, captar más expertos y distribuirles menos trabajo, que viceversa, ya que se obtendrán más valoraciones en menos tiempo.
- El banco de ítems de partida debe ser homogéneo. Para ello es conveniente establecer una guía de estilo de ítems bien para uniformizar un banco existente o bien para crearlo homogeneizado.
- Los ítems del banco han de ser correctos con respecto al área de conocimiento sobre los que versan. Al igual que los ítems para el aprendizaje de idiomas deben cumplir las normas lingüísticas, las de otras áreas deben ajustarse y recoger los avances científicos o nuevas normas y criterios que se establezcan. Esto incluye la revisión periódica de los ítems del propio banco y/o del conjunto de ítems con los que se pretende ampliar, pudiéndose realizar esta revisión mediante un proceso de filtrado.
- En previsión del abandono de participantes, de ítems que puedan quedar sin valoración y de la criba posterior de los datos, hay que fijar el objetivo mínimo de valoraciones por ítem a recoger y

estimar un número superior de valoraciones a completar que garanticen al menos el mínimo preestablecido.

- Es preciso que el análisis y depuración de la muestra se solape y empiece antes de finalizar el proceso de recogida de datos, ya que en tanto en cuanto no se tenga el suficiente número de valoraciones para efectuar una calibración válida, no se podrá dar por finalizada la fase de recogida de datos.
- La calidad de los pronósticos y resultados depende, sobre todo, del cuidado que se ponga en la elaboración del cuestionario y en la elección de los expertos consultados (Landeta, 1999; Stackman, 1974). En el caso de la calibración de ítems, ello implica también la adecuada selección de las variables de calibración.

Capítulo 7
Calibración de ítems con 3PL-TRI (CT)

El proceso de calibración siguiendo el modelo logístico de tres parámetros (3PL) de la Teoría de Respuesta al Ítem (TRI) se ha denominado CT. Para llevarlo a cabo fue necesario conseguir la administración de los ítems a una serie de sujetos. Esta tarea se realizó con dos experimentos desarrollados mediante pruebas de campo, en las que la administración de los sujetos se hizo utilizando dos técnicas diferentes: con supervisión directa de un participante activo y sin ella. Dichas pruebas de campo se llevaron a cabo de manera simultánea y se emplearon para recabar entre ambas al menos 500 valoraciones por ítem. Se ha denominado a cada una de las pruebas como Prueba con TRI 1 (abreviado PT1) y Prueba con TRI 2 (PT2).

Seguidamente se utilizaron los datos conjuntos de ambas pruebas para analizar y realizar la calibración propiamente dicha de los ítems. En las siguientes secciones se presentan cada una de las pruebas realizadas.

7.1. Prueba con TRI 1 (PT1)

El objetivo de la prueba era obtener 250 administraciones de cada uno de los ítems del banco. Se trataba de la mitad de las 500 que (Bunderson, Inouye et al., 1989),entre otros autores, consideran necesarias para calibrar los 3 parámetros de los ítems.

Se crearon 6 cuestionarios electrónicos con 60 ítems cada uno, de los cuales 22 eran de anclaje. En este caso se precisaba de anclaje

puesto que sería necesaria realizar una tarea de equiparación de los valores de dificultad que se asignasen a los ítems ya que a cada sujeto solo se le administraría parte del banco.

Los **cuestionarios** se administrarían **por ordenador en sesiones no supervisadas**. Los cuestionarios estarían disponibles las 24 horas del día en un servidor web para ser respondidos por todos aquellos que lo deseasen. **Se podrían omitir respuestas**, ya que no se obligaría a contestar a todos y cada uno de los ítems del cuestionario, aunque sí se les recomendaría hacerlo. A cambio, una vez completado el cuestionario, se les daría una estimación del nivel de euskera que tuvieran, dando la **sesión** por **completada**. Si el voluntario no llegase hasta este punto de la administración, se consideraría que la **sesión** estaba **incompleta** (inacabada) por abandono. Los cuestionarios debían poderse contestar en a lo sumo 45-50 minutos, si bien el tiempo medio de completado se estimaba iba a ser de 20-30 minutos.

7.1.1. Participantes

Los **sujetos** participantes **activos** fueron un coordinador y ejecutor principal, un supervisor general del proceso y 2 colaboradores para las pruebas piloto y para la puesta en marcha de la administración de cuestionarios vía web.

Los **sujetos pasivos** serían al menos 1500 voluntarios anónimos con diferentes conocimientos de euskara y de informática. Puesto que se iban a emplear los ítems ya corregidos de las pruebas PE1 y PE2, no era necesario revisar la corrección de los mismos.

Se realizarían pruebas piloto durante 2 semanas, intentando captar el mayor volumen posible de participantes. Entre los **revisores** habría especialistas en áreas de psicometría, interacción persona-computador y euskera, que ayudarían a mejorar la versión beta de la herramienta de apoyo a la administración de subtest, además de usuarios con diferente alfabetización informática.

7.1.2. Metodología

Se siguió el siguiente método:

Primero se adaptó **el diseño de los cuestionarios** de las pruebas con expertos para que fueran administrables por ordenador y sirvieran para recabar los datos requeridos para realizar una calibración TRI. Además, se efectuó un nuevo reparto de **los ítems**, de manera que **se distribuyeron homogéneamente en 6 subtests** que compartían un mismo subconjunto de ítems de anclaje. Sendos aspectos se tratan en el apartado 7.1.3.

Paralelamente **se construyó** la versión beta de **la herramienta de administración de cuestionarios electrónicos** (López-Cuadrado, Armendariz et al., 2005). Además de las páginas activas del servidor y que componían los cuestionarios electrónicos, se definió la estructura de la base de datos que albergaría la información relativa a la aplicación de los subtests. La información y relaciones de la base de datos se puede consultar en (López-Cuadrado, 2008). La herramienta de administración permitiría ahorrar muchos problemas de logística y planificación.

Para **captar los revisores** de los cuestionarios electrónicos y de la herramienta de apoyo para la administración de cuestionarios se contactaría directamente con personas del entorno de trabajo del grupo de investigación y que cumplieran con el perfil requerido.

Se realizarían **pruebas piloto** para afinar la aplicación de administración de los tests electrónicos y recabar información, principalmente, para ajustar la estimación del rango temporal válido de las administraciones acabadas y afinar los **criterios de validación** y **criterios de descarte** de sesiones acabadas. El apartado 7.1.4 describe las pruebas piloto desarrolladas y los resultados derivados de las mismas.

Para la **captación de** los sujetos **administrados**, primeramente se contactaría con amistades y compañeros del grupo de investigación, invitándoles a que realizasen un cuestionario electrónico a través de Internet. Adicionalmente, se **enviarían invitaciones de partición abiertas** a listas de distribución de correo electrónico relacionadas con el mundo universitario, el deporte, la cultura y la lengua vasca. En el anexo A4 se puede encontrar algún ejemplo de este tipo de comunicaciones.

Durante el desarrollo de la **administración de los cuestionarios**, la herramienta informática ofrecería al azar a los voluntarios que

desearan participar en la PT1 uno de los 6 cuestionarios, intentando equilibrar los ya realizados. Para ello, a medida que los sujetos anónimos fueran contestando los subtests, se verificaría primeramente si la sesión acabada superaba o no los *criterios de descarte* (véase Tabla 19). Si no se descartaba, seguidamente se contactaría con el sujeto anónimo que la había realizado agradeciendo su participación y solicitándole que indicase explícitamente si su contribución debía incluirse en el proceso estadístico de calibración de ítems o por el contrario debía ser descartada por no haber sido rellenada de forma comprometida para lo que se le enunciarían los *criterios de no validación* de sesión ya fijados (véase Tabla 20). El contacto con el voluntario se establecería empleando el código de identificación de la sesión que éste hubiera suministrado, de manera que si no se conseguía contactar con el mismo, la sesión se consideraría inválida.

Del mismo modo que en las pruebas con expertos, se haría un **seguimiento completo** de los **costes invertidos** y **aspectos mejorables** durante el ciclo de vida de la PT1.

7.1.3. Diseño de los cuestionarios

Los ítems se **distribuyeron homogéneamente en** subtests siguiendo un diseño de **grupos no equivalentes de ítems comunes** (Kolen y Brennan, 1995; Olea, Abad et al., 2002), lo que consiste en repartir los ítems en subtests similares entre sí y lo más parecidos en su totalidad, y compartiendo el mismo subconjunto de ítems de anclaje. La longitud de los subtests estaba sujeta a los 20-30 minutos de media establecidos como tiempo necesario para completar un cuestionario.

La **distribución se realizó a partir** de una clasificación de los ítems en destrezas gramaticales y categorías de dificultad construida con una parte de las valoraciones de los expertos de la PE1. Se descartaron los ítems 240 y 252 del banco, porque no evaluaban claramente ninguna de las 6 destrezas contempladas por Hezinet a la hora de determinar el nivel inicial de un nuevo alumno. Se marcaron nuevos ítems, hasta un total de 35, como potencialmente erróneos. La Tabla 18 recoge los ítems marcados como potencialmente erróneos hasta el momento atendiendo al origen de la causa: por

contenido, por categoría de dificultad, por destreza lingüística y, por contenido y destreza. Tanto los algoritmos aplicados como la distribución precisa de los ítems en subconjuntos propios, incluido el de anclaje, puede consultarse en (López-Cuadrado y Arruabarrena, 2005).

RAZÓN DEL MARCADO	IDENTIFICADORES DE ÍTEMS
Por contenido	1, 13, 16, 59, 77, 102, 135, 146, 148, 152, 170, 178, 186, 191, 202, 225, 229, 237, 249.
Por categoría de dificultad	200, 218, 234, 243
Por destreza lingüística	2, 50, 58, 70, 98, 125, 168, 169, 181, 188
Por contenido y destreza	198, 242
Potencialmente erróneos	En total: 35
RETIRADOS	240 y 252

Tabla 18.- Ítems marcados como potencialmente erróneos y retirados del banco antes de la distribución en subtests

La **estructura de los cuestionarios** de las pruebas anteriores se adaptó para ser presentada en el ordenador transformándose en **una secuencia de sucesivas pantallas**. Toda la información se presentó en bilingüe (euskera y castellano). La estructura constaba de una *identificación*, la *introducción*, una *recogida de datos personales*, *instrucciones* para completar el test y uno de los 6 *subtests* finalizando con la *puntuación alcanzada*.

La pantalla de **identificación** solicitaba una *clave de identificación de sesión* (que no de acceso al sistema). Dicha clave era una elección del sujeto, aunque se sugería que fuera algún método de contacto con él para verificar la validez de su contribución, como una dirección de correo electrónico o un número de teléfono.

La **introducción** presentaba el objetivo del trabajo y las instrucciones de rellenado del cuestionario ilustradas con ejemplos.

La **recogida de datos personales** del participante obtenía algunos datos del sujeto con fines estadísticos como la edad, sexo, titulación, experiencia con el euskera o localización.

Las **instrucciones** indicaban cómo responder al subtest utilizando 2 ítems de entrenamiento como ejemplo.

La parte central del cuestionario estaba dedicada a administrar un **subconjunto de los ítems a valorar,** el correspondiente a uno de los subtests. El sujeto tenía que responder mediante una elección

entre 4 opciones presentadas, o bien omitir la respuesta. El **subtest** presentaba uno a uno los 60 ítems que componían el subtest.

La prueba finalizaba con una pantalla donde se indicaba la **puntuación alcanzada.** Esta se presentaba de dos maneras: *absoluta,* que indicaba el porcentaje de ítems acertados de los 60 presentados, y *ajustada,* que ofrecía una puntuación en la que cada ítem correcto valía 1 punto y cada respuesta incorrecta restaba 1/3 puntos (el número de opciones incorrectas que hubiera).

El anexo A4 recoge la secuenciación de las pantallas de uno de los 6 cuestionarios electrónicos y los correos electrónicos remitidos tanto para captar voluntarios como para validar y agradecer su participación.

7.1.4. Pruebas piloto

Las pruebas piloto se desarrollaron en dos fases, de las cuales *una* se hizo conjuntamente con la CE (véase la sección 6.1.4). En esta fase se revisaron y corrigieron los ítems *desde un punto de vista didáctico*, y 22 ítems quedaron marcados como potencialmente erróneos.

Durante la **planificación** de la segunda fase **de las pruebas piloto** de la PT1, se estableció que los revisores debían contribuir a mejorar la versión beta de la herramienta de apoyo para la administración de subtests. Para ello, debían acceder al servidor web y contestar el cuestionario presentado por la aplicación. Las respectivas sesiones de los revisores servirían para ajustar la estimación del tiempo necesario para completar un cuestionario, inicialmente estimado en 50 minutos y que luego fue de 20-30. Al igual que en la PE1, se solicitó a los revisores que ofrecieran propuestas de corrección lingüística y de estilo para las instrucciones y, en esta ocasión también, para las pantallas de los cuestionarios. Además, se les requirió que determinasen la idoneidad de la navegación entre las sucesivas pantallas. En estas pruebas participaron casi 300 revisores.

La **ejecución de la segunda fase** permitió detectar algunos errores en la versión beta, hacer algunos ajustes y lograr la versión final de la aplicación de administración de subtests vía web.

Tras recoger la información proporcionada por los sujetos piloto, se efectuaron algunas **modificaciones en la interfaz de usuario**: se redujo la cantidad de texto a mostrar sistemáticamente (incluyendo enlaces con aparte para que sólo quien tuviera interés la leyera), se revisaron y corrigieron los textos bilingües presentándolos a dos columnas cada una en un color y se eliminó la posibilidad de retroceder en toda la sesión.

Durante las pruebas piloto se identificaron **dos problemas menores de programación**, que fueron fáciles de detectar y arreglar (López-Cuadrado, Armendariz et al., 2005). El primero de ellos estaba relacionado con el formato de almacenamiento de la fecha y la hora y sucedía cuando el sujeto comenzaba la sesión en un día y la finalizaba pasada la media noche. El otro error se creyó que se daba por pulsar repetidamente el botón "Jarraitu/Continuar" por lo que la aplicación intentaba a su vez almacenar la misma respuesta en más de una ocasión. Ambos problemas se resolvieron a tiempo, de manera que no causaron mayores contratiempos en la ejecución de la PT1.

Tras las pruebas piloto, la versión final de administración de tests era compatible con navegadores como Internet Explorer para Windows, Konqueror para Linux y el software de Opera para cualquiera (otra) plataforma.

Como conclusión de las pruebas piloto se estableció, por motivos de fiabilidad en las respuestas recogidas y como solución preventiva, que **toda administración de subtest fuera inválida mientras no se demostrase lo contrario**. El hecho de aplicar un subtest a cualquier persona que se conectase a la aplicación podía hacer que las pruebas resultasen adulteradas, básicamente porque se desconocía si su participación había sido *responsable*, en contraposición con las respuestas al azar o sin criterio de los usuarios que hubieran realizado la prueba *para ver cómo funcionaba*. El **cuestionario** completado debía estar **validado** para que fuera una contribución a considerar en la calibración estadística mediante la **aplicación de criterios específicos**. El proceso de validación de las sesiones acabadas se efectuaría a medida que se fueran registrando las sesiones, sin relegarla a la fase de "análisis y calibración", ya que el proceso era costoso.

Las sesiones ejecutadas durante las pruebas piloto se emplearon para ajustar los criterios de descarte y de no validación de las sesiones. **Los criterios de descarte de sesiones** a emplear se han recogido en la Tabla 19. En concreto, se eliminarían automáticamente los cuestionarios inacabados, y sistemáticamente aquellos con código de identificación de sesión que no fuera un número de teléfono o dirección de correo electrónico. Además, se descartarían los cuestionarios que hubieran sido acabados con excesiva lentitud (superando el tiempo máximo de 50 minutos establecidos), con interrupciones (el tiempo de respuesta a algún ítem fuera superior a 200s) y los respondidos con excesiva celeridad (con tiempo de completado inferior a 5 minutos). En estos casos se consideraba que el sujeto no había participado con suficiente seriedad en la prueba, por lo que su contribución habría podido malograr los resultados de la calibración estadística. La cota superior de 50 minutos para no descartar una sesión se fijó definitivamente atendiendo a los resultados empíricos de las pruebas piloto. En estas pruebas se constató que el 95% de los sujetos piloto habían sido capaces de responder al total de las 60 preguntas planteadas en ese margen de tiempo. Por otro lado, también se descartarían los subtests con todas las respuestas acertadas, al no servir éstas para estimar la dificultad de los ítems. Por último, se descartarían igualmente aquellas sesiones que superando los criterios de descarte anteriormente mencionados, a continuación no hubieran podido ser validadas directamente con sus respectivos autores.

Cuestionario inacabados (abandono sin llegar a la última pantalla)
Código de identificación de sesión erróneo (ni tel. ni email)
Tiempo de respuesta al cuestionario fuera de margen (<5min o >50min)
Sesión entre el 15% más rápidas en ser completada y con habilidad inferior a la media
Sesión con interrupciones: T. de respuesta al ítem fuera de margen establecido (>200s)
Condiciones de administración inadmisibles (falta de interés, falta de atención, interrupciones,…)
Sesión con el 100% de las respuestas acertadas
Sesión con validez no confirmada por autor

Tabla 19.- Criterios de descarte de sesiones completadas

Durante la fase posterior de la recogida de datos en la de "Análisis y calibración de los ítems" se añadió el criterio "Sesiones entre el 15% más rápidas en ser completadas y con habilidad inferior a la media". La idea subyacente era la de invalidar las sesiones ligadas a sujetos que en media hubieran respondido rápido, pero que no

hubieran mostrado un buen nivel de habilidad. Así mismo, el criterio "Condiciones de administración inadmisibles" aunque aparece recogido en este punto de desarrollo, para presentar conjuntamente todos los criterios considerados para validar o no sesiones, en realidad no de detectó hasta la administración de los cuestionarios en la PT2.

Los criterios de no validación serían los que se iban a plantear a los autores de las sesiones para validar o rechazar sus respectivas contribuciones. La Tabla 20 agrupa los tres filtros considerados. En consecuencia, se consideró que las respuestas del administrado serían válidas para el proceso estadístico posterior si había respondido sin consultar a agentes externos y todas las respuestas las había emitido de forma responsable; en caso contrario, su contribución sería eliminada de la muestra.

El sujeto ha realizado consultas (a otros sujetos, apuntes, web,...)
El sujeto ha seleccionado respuestas al azar (sin seriedad, falta de atención, de interés,...)
El sujeto ha realizado la prueba con interrupciones

Tabla 20.- Criterios para no validar sesiones completadas

7.1.5. Resultados

En la administración de los subtests de la PT1 (Tabla 21), que se extendió por dos años, participaron un total de 2017 individuos en sesiones no supervisadas y por iniciativa propia. De éstas, 382 se descartaron de la base de datos automáticamente, al no haber finalizado el subtest administrado. No obstante, de las 1635 sesiones restantes, 660 (40.4%), aun acabadas, se rechazaron. Consecuentemente, finalmente a través de la PT1 se consiguieron 975 sesiones validadas explícitamente por sus autores (59.6% de las acabadas).

MUESTRA	**Sesiones inacabadas**	**Sesiones acabadas**	**Sesiones (acabadas) validadas**	**Sesiones (acabadas) no validadas**
N. sujetos	382	1635	975	660
Tasas con respecto a sesiones concluidas	23.4%		59.6%	40.4%

Tabla 21.-Tamaño de la PT1

En realidad, tal y como muestra Figura 36, del total de las sesiones comenzadas se descartaron el 19% por no haber completado el

cuestionario, y otro 33% por no haber sido validadas. Consecuentemente, únicamente el 48% (975 de 328+1635) del total de las sesiones comenzadas terminaron y se pudieron validar al contactar con su respectivo autor y éste afirmar el rigor de su actuación durante la prueba.

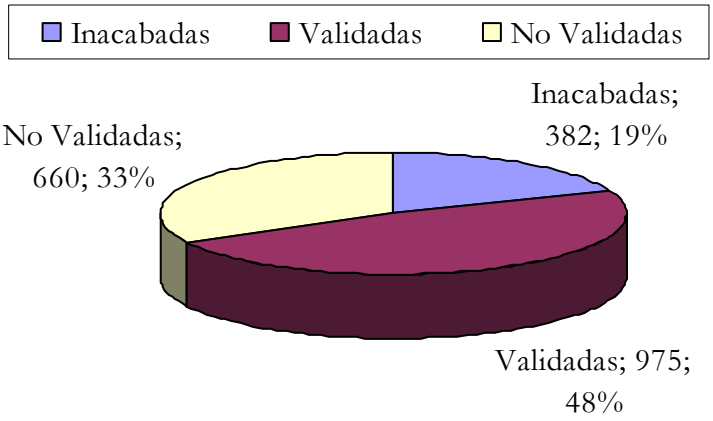


Figura 36.- Tipificación de las sesiones de la PT1

En la Figura 37 se muestra la evolución de las sesiones acabadas desde junio de 2004 a mayo de 2006. En (López-Cuadrado, 2006) se hallan tablas más exhaustivas con los valores de los atributos registrados sobre cada sesión. En términos medios, se recogieron 68 cuestionarios acabados por mes, o bien 41 si se consideran los acabados y validados. Al ser la tasa de sesiones validadas tan pequeña con respecto a las sesiones acabadas (59.6%), para poder alcanzar el objetivo inicial de 1500 cuestionarios validados habría que haber extendido por 13 meses más el desarrollo de la prueba PT1.

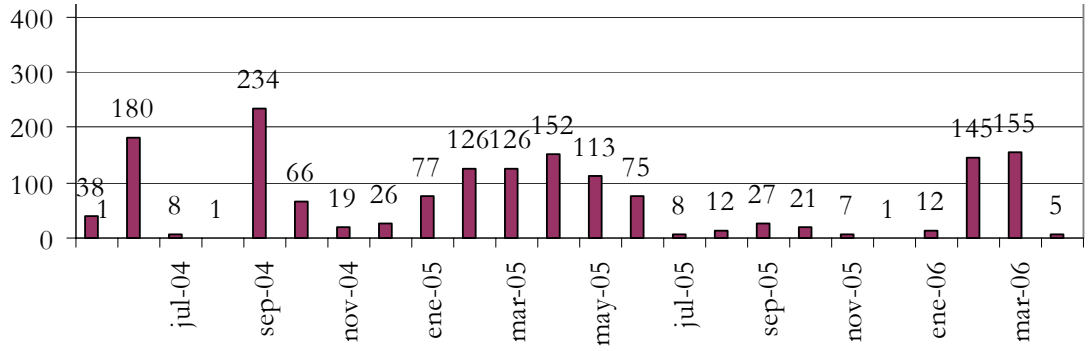


Figura 37.- Subtests acabados en la PT1

En la Tabla 22 se han agrupado el total de las 660 sesiones que se desecharon tras la aplicación de los criterios mencionados en la sección anterior. El motivo principal para desechar las sesiones no supervisadas fue no poder contrastar la validez de la actuación con

su respectivo autor (290 sesiones), seguido de las sesiones realizadas con interrupciones (132) y las que tuvieron código de identificación de sesión erróneo (118), superando estas tres causas más del 80% de los descartes.

Código de identificación de sesión erróneo (ni tel. ni email)	118
Tiempo de respuesta al cuestionario fuera de margen (>50min)	49
Tiempo de respuesta al cuestionario fuera de margen (<5min)	11
Sesión entre el 15% más rápidas en ser completada y con habilidad inferior a la media	30
Sesión con interrupciones: T. de respuesta al ítem fuera de margen establecido (>200s)	132
Condiciones de administración inadmisibles y confirmadas por los autores (falta de interés, falta de atención, interrupciones,…)	25
Sesión con el 100% de las respuestas acertadas	5
Sesión con validez no confirmada por autor (no respondieron al email confirmatorio o no se pudo contactar tel.)	290

Tabla 22.- Desglose de sesiones PT1 accabadas y rechazadas

Restringiéndonos a las 975 se sesiones acabadas y validadas (Figura 38), se recogieron 152 valoraciones en 76 ítems (30%), 164 en 38 (15%), 165 en 38, 171 en 76 y 975 en los 22 ítems de anclaje (9%).

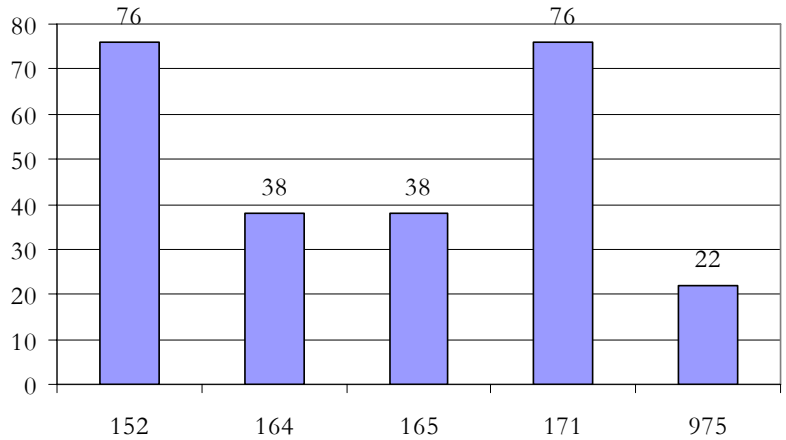


Figura 38.-Valoraciones por ítem recogidas en las sesiones validadas de la PT1

7.1.6. Incidencias

A la vista de la lenta progresión de cuestionarios completados y validados, y tras ejecutar varias medidas para incrementar dicho volumen, se constató que la prueba se dilataría mucho en el tiempo. Por ello se decidió acelerar la puesta en marcha de la PT2 con el objetivo de, una vez alcanzadas las 500 valoraciones entre ambas

pruebas se dieran por concluidas las dos. Como resultado de esta acción, la PT1 se dio por finalizada tras recabar, en términos medios, 162 valoraciones por ítem (el 65% de las 250 previstas inicialmente).

7.1.7. Mejoras

Para una nueva réplica del experimento PT1, se considera **conveniente enviar el grueso de emails de captación simultáneamente** y ampliar el abanico de direcciones a las que se envía la invitación, **incluyendo listas de distribución de empresas, organismos e instituciones que requieren el conocimiento de euskera a sus empleados**, más aún cuando según nuestra experiencia tan solo el 48% de las sesiones comenzadas, acaba y se valida.

Para estimular su participación, a modo de acicate, **se recomienda añadir** a la puntuación absoluta y a la ajustada**, otra puntuación equivalente al nivel de HABE, EOI o IVAP (PL) alcanzado.**

7.1.8. Evaluación de costes

En la Figura 39 se muestra el **tiempo invertido** en el desarrollo de la prueba tanto por los participantes activos como por los pasivos. Se invirtieron 158h en *formación* para realizar la PT1. Las tareas de *planificación y gestión* supusieron 88.8h; ello incluía los tiempos de planificación general de la prueba, y específicamente del filtrado inicial, del anclaje y de la herramienta de administración de subtests, así como el tiempo invertido en el envío de emails de captación de participantes pasivos. Las horas de *implementación* ascendieron a 351.5h lo que incluía el filtrado inicial, el anclaje y el desarrollo de la aplicación web con su correspondiente ajuste, además de la conducción de la prueba y el descarte y validación de las sesiones. Las 777.2h de los sujetos *pasivos* correspondían a las pruebas piloto, el tiempo invertido para completar 1635 subtests y notificar el carácter responsable o no de la participación de los administrados PT1. En términos medios, un sujeto pasivo PT1 necesitó 22.11min para completar un subtest. Este volumen de horas suponía el 52% del número total de horas invertidas en la prueba. Finalmente, la

elaboración de los *entregables* de la fase de recogida de datos de la PT1 requirió 140h. Los 5 aspectos temporales considerados ascendieron a un total de 1515.5h para ejecutar la PT1.

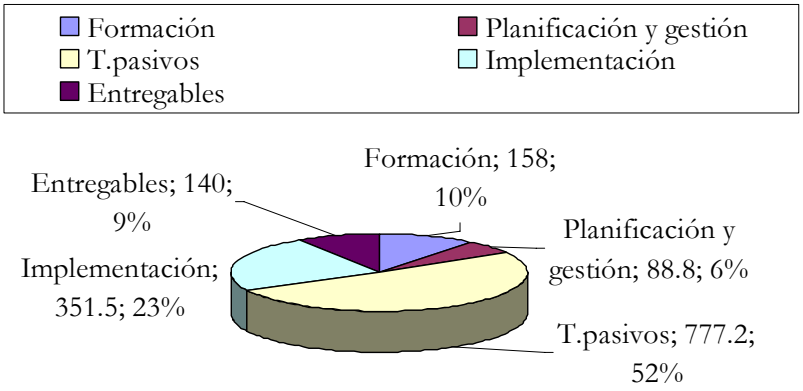


Figura 39.-Tiempo invertido en el desarrollo de la PT1

Además, hubo costes correspondientes a **llamadas telefónicas y correo electrónico** y que se muestran en la Tabla 23. Se estimó que se habían enviado escalonadamente a lo largo del periodo que duró la prueba unos 250 emails a listas de distribución y de particulares, empleando en ello 6.25h. Una vez aplicados los criterios de descarte, y para determinar la posible valía de las 1320 sesiones acabadas (y aún no rechazadas), se enviaron 1032 emails y se realizaron 284 llamadas telefónicas, invirtiendo 34.4h y 18.93h respectivamente. En términos medios, para lograr un cuestionario acabado hubo que enviar 0.15 emails, y para consultar directamente su validez con el autor se invirtieron 3.16 minutos, ya que el monto real de tiempo empleado por los sujetos activos descartando y validando sesiones PT1 ascendió a 86.03h computadas en el apartado *implementación* de la Figura 39.

LLAMADAS TELEFÓNICAS & EMAILS	**N.**	**T. (h)**	**Por cuest. acabado**
Emails captación	250	6.25	0.15 emails
Emails para validar sesiones acabadas	1032	34.4	
Llamadas tel. para validar sesiones acabadas	284	18.93	

Tabla 23.-Consumo de tiempo en llamadas y correos electrónicos en la PT1

A estos costes, hubo que añadirles los costes de un ordenador personal con software de ofimática, Visual Basic y varios navegadores en él instalados a modo de prueba. No hay que olvidar tampoco que se precisó de un servidor web conectado

ininterrumpidamente durante 2 años con su respectiva licencia software de Internet Information Server (Tabla 24).

Recursos económicos
Ordenador personal
Sw. de ofimática
Visual Basic
Servidor web
Sw. Internet Information Server
Conexión de internet

Tabla 24.- Recursos empleados en la PT1

7.2. Prueba con TRI 2 (PT2)

El objetivo de esta prueba de campo era recabar el número de valoraciones necesario para alcanzar, junto con las obtenidas con la PT1, un total de al menos 500 valoraciones validadas por ítem. Se emplearon los mismos cuestionarios de la prueba anterior, aunque en esta ocasión las aportaciones se obtendrían a través de sesiones supervisadas con alumnos en laboratorios con conexión web de centros docentes que acordasen participar en el estudio.

Era necesario llegar a las 3000 participaciones validadas (6 cuestionarios por 500 valoraciones por ítem) para alcanzar entre las ambas pruebas las 500 valoraciones por ítem y poder realizar una calibración según el modelo 3PL.

7.2.1. Participantes

Como **sujetos** participantes **activos**, se identificaron además del coordinador y ejecutor principal, varios colaboradores que principalmente ayudarían a contactar con los centros docentes voluntarios y supervisarían el desarrollo de las administraciones de las pruebas en los laboratorios de los centros. Había otra figura relevante, la del *coordinador de centro docente* donde se realizarían las administraciones. Esta persona haría de nexo entre los responsables de la prueba de campo y los alumnos administrados, colaborando en la distribución de los mismos entre los laboratorios disponibles del centro docente.

Como participantes **pasivos** principales iba a haber una muestra de sujetos anónimos, que serían alumnos de los centros

participantes. También participarían revisores de los cuestionarios electrónicos y de la aplicación informática desarrollada para su administración; si bien no sería necesario disponer de revisores de los ítems, ya que seguían siendo válidas las revisiones realizadas para las pruebas de campo con expertos.

7.2.2. Metodología

La prueba de campo PT2 era similar a la PT1, siendo parejo el proceso de recogida de los datos. Seguidamente se describen las principales diferencias.

Se emplearán los **mismos cuestionarios electrónicos** de la PT1 pero **en sesiones supervisadas** en laboratorios de centros docentes preuniversitarios y universitarios. La herramienta de administración de cuestionarios web generaba al azar ejemplares de subtests intentando equilibrar los ya completados.

Puesto que en términos medios los cuestionarios aceptados de la PT1 estaban precisando unos 30 minutos, únicamente considerando el tiempo de respuesta al subtest, se estableció que **cada sesión de laboratorio durara 30 minutos**.

Únicamente los individuos interesados realizarían la prueba en un laboratorio de su propio centro docente ante un supervisor, por lo que se asumió que **toda administración de subtest era válida salvo que existiera evidencia indicando lo contrario**.

La extensión de la prueba se prolongaría hasta que se obtuvieran los cuestionarios validados necesarios para realizar la calibración 3PL. Para motivar la participación de los mismos, se elaborarían informes para las direcciones de los centros donde se les indicaría los niveles de conocimiento de euskera estimados de los sujetos participantes.

En cada sesión de laboratorio un participante activo presentaría los objetivos de estudio a realizar, indicaría cómo completar el cuestionario y animaría a los sujetos para que lo respondieran de forma responsable. Cada sujeto recibiría un código de identificación de sesión. El participante activo supervisaría la evolución de la realización de las sesiones de laboratorio y ayudaría si surgiera alguna incidencia durante el desarrollo, como problemas de

conexión o los sujetos tuvieran dudas sobre la forma en la que debieran de responder las cuestiones. Sólo invalidaría aquellas sesiones que según su criterio no se hubieran realizado de forma adecuada.

Del mismo modo que en la PT1, se hizo un **seguimiento** exhaustivo de los **costes invertidos** y **aspectos mejorables** durante el ciclo de vida de la PT2.

7.2.3. Resultados

La administración de la prueba PT2 se extendió durante un periodo de 20 meses y todos los subtests iniciados en sesiones supervisadas de laboratorio desarrolladas durante la prueba terminaron (Tabla 25). Sin embargo, aunque no hubo ninguna sesión incompleta, de las 2343 sesiones completadas 75 (3.2%) se rechazaron y 2268 (96.8%) se validaron. Entre ambas pruebas de campo se superó el umbral prefijado de 3000 participaciones válidas.

MUESTRA	Sesiones incompletas	Sesiones concluidas (sin validar)	Sesiones validadas	Sesiones rechazadas
N. sujetos	0	2343	2268	75
Tasas con respecto sesiones concluidas			96.8%	3.2%

Tabla 25.-Tamaño de la PT2

Los 2343 subtests se completaron en 136 laboratorios supervisados, habiendo participado 7 centros preuniversitarios y 3 universitarios. Los 101 laboratorios en los centros preuniversitarios se realizaron en 21 días y gracias a ellos se consiguieron 1693 subtests validados. Mediante los 25 grupos de los centros universitarios se recopilaron otros 575 subtets más. Los primeros centros participaron en otoño del 2004 y el último lo hizo en mayo de 2006. (López-Cuadrado, Arruabarrena et al., 2005) es uno de los 7 informes que se elaboraron durante esta prueba con los perfiles lingüísticos de los sujetos estimados y que se remitieron a sus respectivos centros preuniversitarios.

En la Figura 40 se muestra la evolución de las sesiones acabadas desde finales de 2004 hasta mayo de 2006. En ella se aprecia que, si bien inicialmente se supervisaron unos pocos grupos de laboratorio,

tras el paro de finales del curso 04-05 y tras el verano del 2005, el esfuerzo principal para recabar test mediante sesiones supervisadas se realizó principalmente entre septiembre de 2005 y mayo de 2006.

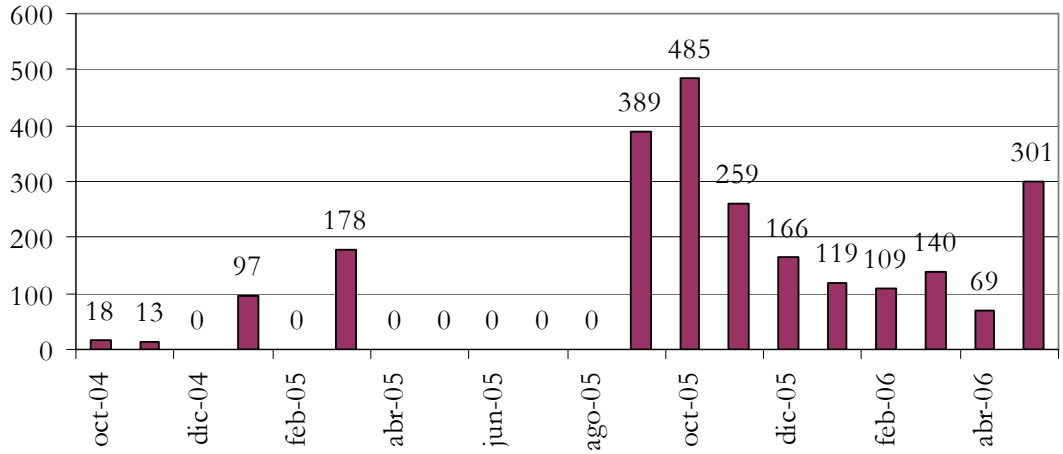


Figura 40.-Subtests acabados en las sesiones de PT2

En la Tabla 26 se han agrupado el total de las 75 sesiones acabadas que se desecharon o bien por aplicación de los criterios de descarte o bien por indicación expresa del supervisor del laboratorio correspondiente. El motivo principal para desechar sesiones supervisadas fue haber estado entre el 15% más rápidas en ser completadas siendo un sujeto con habilidad inferior a la media (51 sesiones), seguido, por igual, de las sesiones con respuestas a ítems superiores a 200s y de las sesiones invalidadas directamente por el supervisor en el laboratorio. Por último, las sesiones descartadas por duración fuera del margen establecido fueron únicamente 4.

Tiempo de respuesta al cuestionario fuera de margen (>50min)	1
Tiempo de respuesta al cuestionario fuera de margen (<5min)	3
Sesión entre el 15% más rápidas en ser completada y con habilidad inferior a la media	51
Sesión con interrupciones: T. de respuesta al ítem fuera de margen establecido (>200s)	10
Condiciones de administración inadmisibles y confirmadas por los supervisores (falta de interés, falta de atención, interrupciones,…)	10
Sesión con el 100% de las respuestas acertadas	0

Tabla 26.- Desglose de sesiones PT2 acabadas pero rechazadas

Restringiéndonos a las 2268 sesiones PT2 acabadas y validadas, en términos de valoraciones de ítems (Figura 41), se recogieron 366 valoraciones en 38 ítems (15%), 374 en 76 (30%), 377, 387 y 390 en otros 3 bloques de 38 ítems, y 2268 en los 22 ítems de anclaje (9%). En términos medios se recabaron 378 valoraciones por ítem, sin contabilizar los de anclaje.

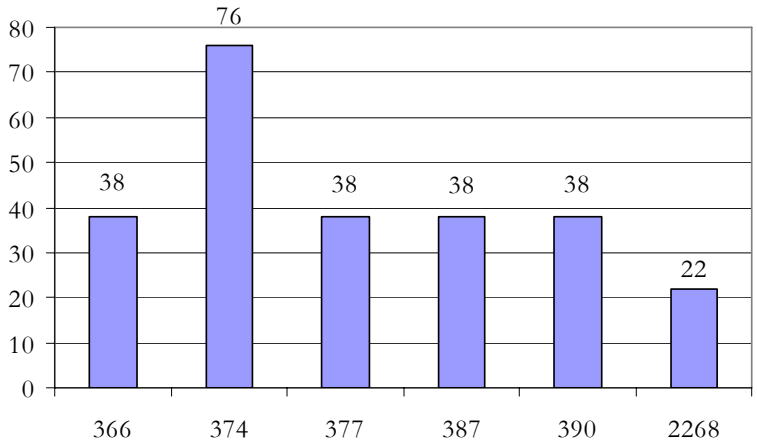


Figura 41.- Valoraciones por ítem recogidas en la PT2

7.2.4. Incidencias

No hubo incidencias destacables, sólo incidentes menores que no afectaron al resultado de la prueba de campo.

7.2.5. Mejoras

En futuras réplicas de pruebas PT2, y para no dilatar la extensión en el tiempo de la recogida de datos, se propone **concertar con los centros docentes las sesiones de laboratorio durante un periodo compacto de tiempo**.

Se ha constatado que la inversión de tiempo en planificar redunda en un mayor número de participantes pasivos conseguidos. Para fomentar la participación de estos, conviene promover que los responsables de los grupos candidatos a participar en el estudio sean quienes establezcan **las fechas** de la celebración de las sesiones de laboratorio, con el objetivo de que éstas **no interfieran con sus actividades docentes** y resulten, por tanto, **lo más oportunas posibles**. Si bien esto puede ocasionar que se tenga que visitar el centro en más de una ocasión, el inconveniente se contrarresta por el incremento del volumen de encuestados, que a su vez favorece el desarrollo de la etapa de recogida de datos y hace que su fin esté más próximo.

7.2.6. Evaluación de costes

El **tiempo invertido** por los participantes activos y pasivos se ha recogido en la Figura 42. El tiempo invertido en la PT2 en *formación* ascendió a 158h, el mismo tiempo que en la PT1, ya que ambas pruebas requirieron de la misma formación. Las tareas de *planificación y gestión* precisaron 98.5h; lo que incluía los tiempos de planificación general de la prueba, y en concreto del filtrado inicial, del anclaje y de la herramienta de administración de subtests, también incluía el tiempo de captación de centros y organización de sus laboratorios. El coste de horas de los sujetos activos durante la *implementación* de la PT2 ascendió a 607.6h; distribuidas entre el desarrollo del filtrado, del anclaje, de la aplicación web, las horas invertidas en los laboratorios, en el desplazamiento de los sujetos activos y el estudio de descarte de sesiones. Debido al volumen de la muestra, la elaboración de *entregables* fue de 150h en la PT2. Por último, hay que indicar que el volumen de 753.4h atribuibles a los sujetos *pasivos* incluía las horas realizando las pruebas piloto y el rellenado de los 2343 subtests. En término medios, el administrado PT2 invirtió 18.91min completando el subtest. El tiempo invertido por los coordinadores de los 7 centros preuniversitarios (a razón de 2h/coordinador) no se acumuló al tiempo total de los participantes pasivos. Los 5 aspectos temporales considerados sumaron un total de 1767.5h para ejecutar la PT2.

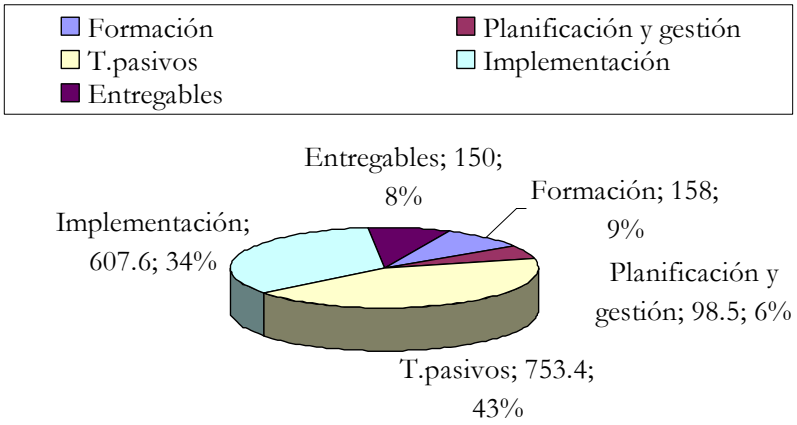


Figura 42.-Tiempo invertido en el desarrollo de la PT2

Desde el punto de vista de **costes** asociados a **llamadas telefónicas**, en la PT2 para formalizar la captación y gestionar con el coordinador del centro preuniversitario los días de las sesiones así como la distribución de los alumnos en laboratorios, se realizaron

unas 70 llamadas (10 llamadas/centro preuniv.) con 10.5h de duración total. Desde el punto de vista de cuestionario, significó que para obtener un cuestionario completado se precisaron 0.27 minutos de llamadas telefónicas.

LLAMADAS TELEFÓNICAS	**N.**	**T. (h)**	**Por cuest. acabado**
Llamadas tel. con centros preuniversitarios	70	10.5	0.27min

Tabla 27.-Consumo en llamadas y correos electrónicos en la PT2

La prueba PT2 requirió 21 **desplazamientos** de ida y vuelta a los centros (Tabla 28). Consecuentemente, los participantes activos invirtieron 55.17h para desplazarse un total de 1556km. Esto supuso que se desplazaron 0.66km para obtener un cuestionario rellenado. Se pueden consultar las tablas de (Arruabarrena, López-Cuadrado et al., 2007) para mayor detalle de los costes presentados en esta sección.

DESPLAZAMIENTOS	**En total**	**T. activos (h)**	**Por cuest. acabado**
Km. desplazados hasta centros preuniversitarios	1556	55.17	0.66km

Tabla 28.- Consumo en desplazamientos de la PT2

A estos costes, además, hubo que añadirles los costes de un ordenador personal con licencias de software de ofimática y Visual Basic así como varios navegadores en él instalados a modo de prueba. Al igual que en la PT1, se precisó también de un servidor web con licencia software de Internet Information Server conectado ininterrumpidamente durante 2 años. Los recursos citados se hallan en la Tabla 29:

Recursos económicos
Ordenador personal
Sw. de ofimática
Visual Basic
Servidor web
Sw. Internet Information Server
Conexión a Internet
Medios de transporte

Tabla 29.- Recursos empleados en la PT1

7.3. Análisis y calibración con TRI

A la hora de realizar la calibración estadística, la muestra constaba de 3243 subtests completados y validados por otros tantos sujetos anónimos. Cada sujeto había contestado a 60 ítems; en promedio, por ítem de anclaje se habían alcanzado 3228 respuestas y por el resto de ítems 540. Esta cifra era superior al mínimo de 500 valoraciones establecidas como objetivo a recoger para realizar una calibración empleando el modelo de logístico de tres parámetros (3PL) de la Teoría de Respuesta al Ítem (TRI).

Si bien de dicha muestra de respuestas ya se habían descartado administraciones inaceptables (como sesiones inacabas, acabadas con demasiada celeridad o por extenderse en exceso en el tiempo), era conveniente realizar otros **análisis previos para depurar** todavía más **la muestra**, antes de proceder con el proceso de **obtención de los** valores de los **parámetros**. Los 4 análisis específicos realizados se han recogido en la sección 7.3.1, mientras que la obtención de los valores de los parámetros se halla en la sección 7.3.2.

El desarrollo principal de la fase análisis y calibración estadística se llevó a cabo durante la primavera-otoño de 2006.

7.3.1. Depuración de la muestra

En este momento el volumen de la muestra ascendía a 194580 respuestas correspondientes a 3243 administraciones acabadas y validadas.

La etapa principal de depuración de la muestra comenzó con un análisis de **protocolos de respuesta anómalos**. El análisis serviría para detectar respuestas anómalas e ítems defectuosos, que seguidamente se retirarían del proceso de calibración.

La administración electrónica de los subtests evitó la existencia en la entrada de datos de respuestas con selección múltiple marcada o bien con selección fuera de rango. Sin embargo, hubo que estudiar las *omisiones de respuestas* registradas, ya que el sistema no obligaba al examinado a responder las preguntas. Respecto a los porcentajes de omisiones, se comprobó que durante las administraciones de los

subtests se dejó sin responder un promedio del 0.55% de los ítems presentados a los examinados. El subtest con mayor tasa de respuestas omitidas en promedio fue el subtest 1 con un porcentaje del 0.88. Por su parte, cada ítem fue omitido en media el 0.31% de las veces que fue administrado, siendo el máximo porcentaje un 3.9 (correspondiente al ítem 180). En todos los casos se trató de valores tan pequeños que las omisiones en las administraciones se pudieron considerar como respuestas erróneas sin que ello pudiera suponer ningún inconveniente durante la estimación de los parámetros (Olea y Ponsoda, 2003).

Seguidamente se buscaron *patrones de respuesta anómalos por administración*, como administraciones con la misma opción de respuesta seleccionada siempre o en las que no se hubiera seleccionado al menos una vez cada opción en ítems diferentes. La búsqueda resultó infructuosa, ya que no se detectó ninguna sesión de estas características.

A continuación se buscaron *patrones de respuesta anómalos por ítem*, como ítems con distractores u opciones de respuesta que no se hubieran seleccionado nunca o porcentajes de respuesta de cada uno de los distractores. Se identificaron 13 ítems (con identificadores 1, 3, 4, 9, 14, 24, 34, 37, 53, 87, 97, 103 y 124) en los que un distractor no se había seleccionado, y un ítem (con identificador 21) con dos distractores intactos. Pese a que en principio este hecho no era significativamente grave, los 14 ítems citados se marcaron como potencialmente erróneos.

Posteriormente se procedió con el **análisis clásico de fiabilidad**, mediante el cual se descartarían ítems con características extremadamente desfavorables. Concretamente, para efectuar los estudios de fiabilidad para los 22 ítems del conjunto de anclaje se recurrió al paquete estadístico SPSS. Se obtuvieron indicadores de fiabilidad aceptables (en concreto, una alfa de Cronbach y un coeficiente de Spearman-Brown ligeramente superiores a 0.8) y se identificaron 6 ítems defectuosos, por tener correlación ítem-subtest muy inferior a 0.3 (con identificadores 5, 6, 129, 189, 205 y 227), que se retiraron del banco. Se repitió el análisis para los 16 ítems de anclaje restantes, obteniéndose para las 3243 sesiones de la muestra unos índices de fiabilidad que superaban en un par de centésimas los

logrados anteriormente. La correlación elemento-total obtenida en el nuevo análisis resultó para cada uno de los 16 ítems del conjunto de anclaje superior a 0.3.

Posteriormente se procedió a realizar estudios de fiabilidad para los ítems de cada uno de los seis subtests. Para todos ellos se obtuvieron alfas de Cronbach y coeficientes de Spearman-Brown superiores a 0.8, por lo que los resultados resultaron aceptables. En lo referente a la eliminación del banco de ítems de aquellos que no correlacionaban adecuadamente con el total del subtest, durante el análisis clásico de fiabilidad para los seis subtest se definió un punto de corte para la correlación elemento-total de 0.25. En consecuencia, se retiraron 40 ítems cuya correlación con el subtest resultó menor que este umbral (Tabla 30). Se marcaron como potencialmente erróneos 25 ítems con valores de correlación elemento-total superiores a 0.25 e inferiores a 0.3. Los valores de estos 25 ítems estaban muy cercanos al valor de corte 0.3 empleado para los ítems de anclaje.

Una vez eliminados los 40 ítems con correlación elemento-total inferior a 0.25, se repitió el análisis de fiabilidad clásica para cada uno de los seis subtests. Para todos los subtests se obtuvieron alfas de Cronbach y coeficientes de Spearman-Brown ligeramente superiores a los que se disponía de antes. Los valores concretos de los coeficientes considerados pueden consultarse en (López-Cuadrado y Armendariz, 2006).

Nº de subtest	**Nº de ítems retirados**	**Ítems retirados**
Anclaje	6	5, 6, 129, 189, 205, 227
1	4	4, 16, 18, 200
2	5	1, 11, 121, 159, 213
3	7	2, 25, 57, 135, 161, 168, 234
4	5	12, 14, 151, 239, 250
5	9	21, 56, 82, 145, 173, 182, 199, 226, 249
6	10	3, 23, 26, 123, 148, 153, 191, 225, 236, 238

Tabla 30.- Ítems descartados durante el análisis clásico de fiabilidad

El siguiente análisis realizado verificó la propiedad de **unidimensionalidad** de los 204 ítems que prevalecían en el banco, empleando la submuestra obtenida de eliminar de la muestra de partida las respuestas de los ítems retirados. Una vez más, se realizó primeramente el análisis para los 16 ítems del conjunto de anclaje y,

a continuación, para los ítems de los 6 subtests (incluidos los anteriores).

Por disponer de una muestra de respuestas lo suficientemente grande para *los ítems del conjunto de anclaje*, se procedió a efectuar un análisis *factorial exploratorio y otro confirmatorio*. La matriz de correlaciones tetracóricas se obtuvo mediante el preprocesador PRELIS y el índice de Kaiser-Meyer-Olkin, un estadístico que examina si las correlaciones parciales entre las variables son pequeñas, también auguró buenos resultados en el análisis factorial, por ser 0.941, un valor superior a 0.7, que es el mínimo aceptable (Kaiser, 1974). Durante el análisis factorial confirmatorio, los índices que proporcionó el software LISREL fueron aceptables, por lo que se determinó que los 16 ítems de anclaje medían un único rasgo simultáneamente y satisfacían, por tanto, el supuesto de la unidimensionalidad.

Para estudiar la *unidimensionalidad de los ítems* de cada uno *de los subtests* se disponía de una muestra menos amplia que para la del conjunto de anclaje, de ahí que se realizara un *análisis factorial exploratorio*, y no confirmatorio, mediante la combinación del paquete estadístico SPSS 13.0 para Windows y el tándem PRELIS/LISREL. Para todos los subtests los índices Lumsden (es decir, el cociente entre el primer y el segundo factor) obtenidos fueron superiores a 5. Por tanto, también aquí pudo concluirse que los ítems de los seis subtest satisfacían el supuesto de unidimensionalidad.

Para finalizar con los estudios previos a la estimación de los parámetros, se ejecutó un **análisis del funcionamiento diferencial de los ítems (FDI)** para descartar ítems sesgados. El estudio consistió en comparar las diferencias entre los resultados obtenidos en las sesiones validadas supervisadas y las no supervisadas. Se identificaron diferencias significativas en los porcentajes de respuestas registrados en ambos tipos de sesiones para un total de 51 ítems, que pueden consultarse en (López-Cuadrado y Armendariz, 2006). No obstante, 19 de ellos (los de identificador 3, 14, 18, 21, 26, 121, 123, 135, 148, 153, 159, 168, 173, 191, 200, 213, 225, 239 y 250) correspondían a ítems que ya habían sido eliminados del banco tras los análisis de fiabilidad, y otros 16 (con

identificadores 13, 19, 24, 66, 73, 125, 155, 186, 188, 207, 208, 218, 221, 242, 243 y 251) se encontraban ya marcados como potencialmente erróneos (López-Cuadrado y Arruabarrena, 2005).

Como **resumen de esta etapa** previa a la obtención de los parámetros, hay que subrayar que los diversos análisis efectuados sobre los ítems y los datos brutos fueron mermando tanto el volumen del banco de ítems como el tamaño de la muestra considerada, evolución que se ha recogido en la Tabla 31. Así, aunque el banco original constaba de 252 ítems, antes de la administración de los subtests ya se habían descartado 2 ítems (campo *"Inicio"*). Con la administración de los subtests se obtuvieron 193630 respuestas validadas para los 250 ítems que quedaban (campo *"Tras administración"*). Sin embargo, como consecuencia de análisis previos a la calibración realizados, se descartaron 48 ítems más junto con sus 40822 respuestas, lo que mermó el volumen de la muestra hasta las 152808 entradas (columna *"Tras depuración"*). En términos porcentuales, los análisis previos mermarón el banco de ítems en un 19% con respecto el banco original y la muestra en un 21.1%.

Tamaño	Inicio	Tras administración	Tras depuración
N. de ítems	252	250	204 (81%)
N. de respuestas		193630	152808 (78,9%)

Tabla 31.- Características de las muestras estadísticas consideradas

Desde el punto de vista de subtests acabados (Tabla 32), se recabaron 3975 cuestionarios, que eran, en promedio, 662.2 contribuciones por subtest o ítem. Tras los análisis previos se había descartado el 18.41% de las aportaciones, validándose 3243 subtests, los cuales contenían una media de 540 valoraciones por ítem.

N. sesiones acabadas	Subtest1	Subtest2	Subtest3	Subtest4	Subtest5	Subtest6	Total
PT1	171	152	171	165	152	164	975
PT2	374	390	366	374	387	377	2268
Rechazadas	101	115	145	104	138	130	732
Total (validadas)	646 (545)	657 (542)	682 (537)	643 (539)	677 (539)	671 (541)	3975 (3243)

Tabla 32.- Resumen de sesiones validadas y rechazadas por subtest

La Tabla 33 recoge los identificadores de los ítems marcados como potencialmente erróneos así como los retirados del banco tras los análisis previos a la estimación de parámetros.

	Ítems	Subtotal
Retirados tras análisis por expertos	240, 252	2
Retirados tras análisis fiabilidad (anclaje)	5, 6, 129, 189, 205, 227	6
Retirados tras análisis fiabilidad (subtest)	1, 2, 3, 4, 11, 12, 14, 16, 18, 21, 23, 25, 26, 56, 57, 82, 121, 123, 135, 145, 148, 151, 153, 159, 161, 168, 173, 182, 191,199, 200, 213, 225, 226, 234, 236, 238, 239, 249, 250	40
RETIRADOS	**Total:**	**48**
Marcados por contenido	59, 77, 102, 152, 170, 178, 229, 237	8
Marcados por destreza lingüística	50, 58, 70, 98, 169	5
Marcado por contenido y destreza	198	1
Marcados por tener opciones nunca elegidas	34, 37, 87, 97, 103, 124	6
Marcados por correlación	36, 46, 52, 107, 112, 165, 187, 195, 224, 230, 246	11
Marcados por diferencias en los porcentajes	138, 210	2
Marcados por funcionamiento diferencial (FDI)	8, 27, 85, 89, 92, 106, 110, 147, 175, 204	10
Marcados por diferencias en los porcentajes y FDI	40, 68, 76, 118, 130, 139, 150, 160, 172, 179, 180, 214, 219, 220, 241	15
Marcados por correlación, porcentajes y FDI	19, 66, 155, 207, 208, 221, 251	7
Marcados por dificultad, porcentajes y FDI	218, 243	2
Marcados por correlación y contenido	146, 202	2
Marcados por correlación y opciones de respuesta	9, 53	2
Marcados por contenido y porcentajes	13, 186	2
Marcado por correlación, destreza, porcentajes y FDI	188	1
Marcado por destreza y porcentajes	125	1
Marcado por destreza lingüística y FDI	181	1
Marcado por opciones de respuesta, porcentajes y FDI	24	1
Marcado por contenido, destreza, porcentajes y FDI	242	1
Marcado por correlación y porcentajes	73	1
Marcado por correlación y FDI	28	1
POTENCIALEMENTE ERRÓNEOS	**Total:**	**80**

Tabla 33.- Ítems descartados y marcados como potencialmente peligrosos tras los análisis previos a la estimación de los parámetros

Tras los análisis previos a la calibración, se retiraron 48 ítems y la muestra depurada contenía 204 ítems, de los cuales 80 estaban etiquetados como potencialmente erróneos.

7.3.2. Calibración de la dificultad

La primera de las tres actividades a realizar para determinar los parámetros de los ítems fue **estimar los parámetros de los mismos** según el modelo logístico de tres parámetros. Para ello se utilizó el método de estimación bayesiana MAP implementado por el software XCALIBRE 1.10 para Windows (ASC, 1997), configurado con un máximo de 12 iteraciones y con distribuciones previas $N(0.75, 0.12)$ para el parámetro de dificultad, $N(0, 1)$ para la discriminación y $N(0.25, 0.025)$ para el pseudoacierto, pudiéndose consultar más detalles de la misma en (López-Cuadrado y Armendariz, 2006).

La Tabla 34 muestra un resumen de los resultados del proceso de estimación de los parámetros para el conjunto de anclaje y cada uno de los seis subtests de los siete conjuntos. Puede observarse que los ítems del banco resultaron ser en general fáciles, pues todas las dificultades medias obtenidas fueron negativas, quedando por debajo del punto medio de la escala (<0.0). En lo que se refiere a los índices de discriminación y pseudoacierto, los valores obtenidos fueron aceptables, ya que los primeros rondaban en término medio el valor 1, y los segundos quedaban por debajo de 0.25, que es el valor probabilísticamente esperado para ítems de cuatro alternativas de respuesta. Unos párrafos más adelante se retomará el análisis sobre el rango de los valores de los parámetros una vez unificadas las puntuaciones.

Subtest	a_i Discriminación (media/desv.típ)	b_i Dificultad (media/desv.típ)	c_i Pseudoacierto (media/desv.típ)	Iteraciones realizadas	Mayor cambio en última iteración	Residual máximo	Índice KR21
Anclaje	0.97 / 0.27	-1.08 / 0.68	0.14 / 0.01	7	0.043	**1.86**	**0.831**
1	1.12 / 0.22	-0.82 / 0.80	0.22 / 0.01	4	0.043	1.44	0.939
2	1.02 / 0.18	-0.99 / 0.92	0.23 / 0.01	4	0.044	1.35	0.921
3	1.00 / 0.20	-1.06 / 1.06	0.21 / 0.01	5	0.028	1.67	0.918
4	1.11 / 0.22	-0.92 / 0.81	0.22 / 0.01	4	0.049	1.38	0.937
5	1.05 / 0.22	-0.76 / 0.92	0.21 / 0.01	5	0.024	1.73	0.919
6	1.02 / 0.19	-0.95 / 0.84	0.22 / 0.01	4	0.039	1.58	0.924

Tabla 34.- Resultado del proceso de estimación de los parámetros de la calibración CT

Respecto al proceso de estimación en sí mismo, pudo comprobarse que la convergencia se produjo en menos de 8 iteraciones frente a las 12 establecidas como máximo, y que el mayor

cambio en los parámetros en la última iteración fue siempre menor que 0.05. Asimismo, el índice de confiabilidad KR21 (Kuder y Richardson, 1937), proporcionado por el propio XCALIBRE, superó siempre el valor mínimo aceptable, que para algunos autores es 0.6 (Lafourcade, 1971) y para otros 0.7 (Kearns, 1998). Por último, los residuales no fueron en ningún caso mayores que 1.96, valor establecido como el máximo permitido para que, con una confianza del 95%, la estimación de parámetros no quedase comprometida. Los valores de los índices obtenidos confirman el **ajuste de los datos al modelo 3PL empleado**.

Como último paso para determinar los valores de los parámetros, se procedió a **equiparar las puntuaciones** obtenidas. Las estimaciones de los parámetros de los ítems obtenidas compartían escalas de medida diferentes, pues se habían obtenido a partir de procesos estadísticos independientes. Gracias a que se hubo definido un diseño de anclaje de ítems antes de proceder a la etapa de administración de los mismos (López-Cuadrado y Arruabarrena, 2005), en este punto fue posible equiparar las métricas utilizadas por los seis subtests, y en consecuencia pudo decirse que el banco de ítems estaba finalmente calibrado.

De entre las diferentes alternativas existentes se optó por recurrir al *método de equiparación media-sigma* (Marco, 1977), por tratarse del más habitual. Además se trata de un procedimiento muy sencillo de implementar, tanto que pudo obtenerse una métrica común a todo el banco con ejecutar una serie de operaciones sobre una hoja de cálculo Microsoft Excel 2000. Los rangos de los valores de los tres parámetros determinados se han recogido en la Tabla 35.

Parámetros	**Rango**
Discriminación (a_i)	[0.6, 1.7]
Dificultad (b_i)	[-2.82, 1.66]
Pseudoacierto (c_i)	[0.11, 0.24]

Tabla 35.- Rango de los valores medios determinados para los 3 parámetros logísticos

Con respecto al parámetro dificultad, los valores estimados de la Tabla 34 ya anticipaban que los ítems del banco resultaban ser en general fáciles al quedar por debajo del punto medio de la escala las medias de dificultad de los subtests. Nótese que éste era un hecho esperado, dado que los ítems del banco estaban pensados para

evaluar a nuevos usuarios del sistema Hezinet, cuyo nivel de euskara era previsiblemente bajo. Figura 43 muestra la curva característica del banco de ítems, de la cual se infiere que la dificultad del banco (como unidad) es próxima al -0.6, luego, de dificultad media-baja. Concretamente, el 71% de los ítems tenía una dificultad estimada inferior a dicho valor (145 ítems de 204 prevalecen en el banco). Añadir que la distribución de las dificultades calculadas del banco de partida no fue ni simétrica ni homogénea, prácticamente careciendo de ítems con dificultad elevada.

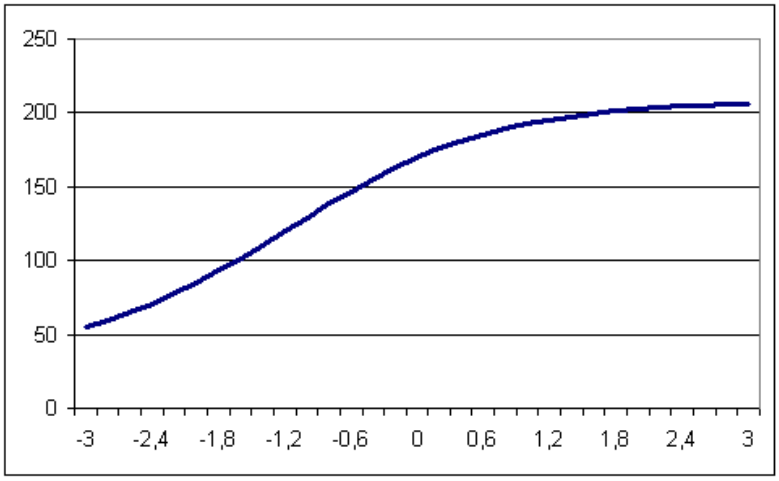


Figura 43.- Curva característica del banco de ítems según CT

El anexo A5 de esta memoria muestra el resultado final de la calibración del banco de ítems según el modelo logístico de tres parámetros de la TRI. En concreto, recoge los valores de discriminación (*a*), dificultad (*b*) y pseudoacierto (*c*) de los 204 ítems que superaron los procesos de filtrado, validez y bondad de ajuste, así como el motivo de eliminación de los 48 ítems que fueron retirados del banco original. Así mismo, en él se destacan los ítems que, pese a permanecer en el banco, fueron marcados como potencialmente erróneos.

7.3.3. Evaluación de costes

La Figura 44 muestra los **tiempos invertidos** para realizar el desarrollo de la segunda fase de la calibración CT. Se calculó que se emplearon 80 horas en formación, 66 en planificación y gestión, 27 en implementación (20 invertidas en "análisis previos" y las otras 7 en "obtención de parámetros), 8 horas imputables al análisis de los

resultados y 80 horas más documentando el proceso. Las horas invertidas en asesoramiento por parte de los dos psicómetras y los dos desarrolladores principales se incluyeron dentro del apartado *planificación y gestión*. La estimación de horas invertidas para realizar la segunda fase de la calibración estadística asciendió a **261** horas.

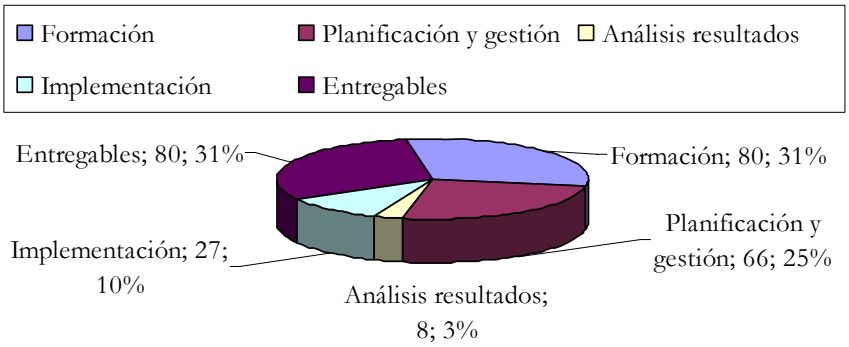


Figura 44.- Estimación del coste de horas invertidas en la fase "Análisis de datos y Calibración" de CT

Desde la perspectiva de recursos económicos (Tabla 36), se dispuso de un ordenador personal con acceso a Internet y con licencias software de un paquete de ofimática, SPSS, LISREL/PRELIS, XCALIBRE. Además, para recibir asesoramiento psicométrico, los investigadores realizaron varios desplazamientos.

Recursos económicos
Ordenador personal y sw. de ofimática
Licencias sw: SPSS LISREL/PRELIS XCALIBRE
Desplazamientos

Tabla 36.- Recursos empleados en la fase "Análisis de datos y Calibración" de CT

7.4. Síntesis

En esta sección se sintetizan los **resultados generales obtenidos** de las fases de "recogida de datos" y de "análisis de datos y calibración" de ítems atendiendo al rasgo Dificultad según el modelo 3PL de la TRI. Seguidamente se presentan las **lecciones aprendidas**. El análisis de los recursos empleados se pospone hasta el siguiente capítulo, en el cual se contrastarán los costes de ambas calibraciones realizadas.

Para **recabar al menos 500 valoraciones por ítem**, se realizaron dos experimentos solapados en el tiempo implementados a través de

sendas pruebas de campo (la PT1 y la PT2). Mientras que en la PT1 los sujetos anónimos completaban un cuestionario en sesiones no supervisadas, en la PT2 lo hacían en sesiones de laboratorio supervisadas. En cada una de las sesiones, el cuestionario administrado contenía un subtest con 60 ítems, de los cuales 22 eran de anclaje. En total, hubo 6 subtests distintos. Con la conducción de ambas pruebas se obtuvieron un total de 3243 subtests acabados y validados, y se invalidaron 735 subtests acabados. Los subtests incompletos se descartaron directamente, y no se consideraron.

El objetivo de la fase de recogida de datos se **cumplió**, ya que, en términos medios, y sin considerar las valoraciones de los ítems de anclaje, el número de valoraciones por ítem obtenidas fue de 540 (el 30% de la PT1 y el 70% de la PT2) y el de los ítems de anclaje de 3227, cifra muy superior. No obstante, señalar que si bien inicialmente se esperaba recabar un volumen similar de valoraciones con ambas pruebas, y habiendo tomado medidas de contingencia durante la administración de la PT1, no se pudo conseguir.

Las tasas de abandono de sesiones **y las tasas de sesiones acabadas** pero **invalidadas** son otros aspectos a destacar. Desde el inicio del proceso global de la calibración estadística se planificó adelantar parte del análisis de la muestra a la fase de recogida de datos. Así, al comienzo de la misma se concretaron los criterios de descarte y de no validación, criterios que se ajustaron tras las pruebas piloto (véanse la Tabla 19 y la Tabla 20 respectivamente). La aplicación de estos criterios descartó un 40.37% de las sesiones acabadas de la PT1 frente a un 3.2% de las sesiones de la PT2 (en promedio el 18.5% de las sesiones acabadas). Además, en la PT1 hubo sesiones en las que los participantes abandonaron la prueba dejando incompleto el cuestionario suministrado (un 23.4% con respecto a las acabas), mientras que en la PT2 toda sesión iniciada se completó.

Una vez obtenida una muestra lo suficientemente amplia para realizar la calibración estadística determinada, se procedió a depurar la muestra con **los análisis previos** identificados en la bibliografía especializada, que **dotaron de validez estadística los resultados obtenidos**. Específicamente se realizó un *análisis de fiabilidad*, comprobado la *unidimensionalidad* de los ítems, se estudió el

funcionamiento diferencial de los ítems y se comprobó la *bondad de ajuste* de los datos al modelo 3PL empleado. Resaltar que los múltiples indicadores, coeficientes e índices computados en dichos análisis obtuvieron valores aceptables.

Como resultado de los análisis previos realizados, se descartaron otros 46 ítems más (además de los 2 descartados en la fase anterior) y quedaron marcados como potencialmente erróneos 80 ítems (recogidos en la Tabla 33). En términos porcentuales, equivalía a **descartar** en total **el 19% de los ítems** del banco y el 21.1% de las aportaciones de los sujetos anónimos.

Seguidamente, se procedió **con la calibración** de los 204 ítems que quedaban en banco. Para ello, primeramente se obtuvieron los valores de los parámetros logísticos mediante aplicación del método de *estimación bayesiana MAP*, y posteriormente se equipararon las puntuaciones mediante el método *media-sigma*. **El rango de los valores medios de los parámetros discriminación, dificultad y psuedoacierto obtenidos fueron** igualmente **aceptables**. Los valores de los parámetros dificultad revelaron que, al igual que en la calibración basada en expertos, los ítems del banco eran, en general, fáciles.

Una de las **lecciones aprendidas** de la puesta en marcha de la pruebas con sujetos anónimos es que para decantarse por una u otra prueba de campo, además de la eficiencia de costes económicos y temporales, hay que sopesar las tasas de abandono de sesiones y de descarte de sesiones acabadas. Esto significaría que para obtener al menos 500 valoraciones con sesiones PT1, se precisarían 5031 sesiones no supervisadas completadas, frente a 3096 sesiones de laboratorio. Considerando los miembros actuales del grupo de investigación, sus respectivos contactos así como su capacidad de persuasión, obtener datos solo a través de sesiones PT1 en un periodo de tiempo aceptable ha sido, y se ve, inviable.

En el supuesto de que en un futuro se decantara por recabar información mediante sesiones PT1, y en cuanto a captación de participantes voluntarios se refiere, según nuestra experiencia (1635 participantes completan un test de 60 preguntas), hay dos aspectos primordiales a tener en cuenta: la carta de invitación y los foros de divulgación de la misma. La carta de invitación a participar en las

pruebas de campo tipo PT1 debe redactarse con sumo cuidado. Igualmente, se debe poner especial atención a la hora de escoger los foros o listas de distribución adecuados a los que se remitirá ésta. No obstante, nótese que, si bien estos dos avales ayudan a obtener un volumen de participación positivo, una distribución exponencial de invitaciones no implica una participación de voluntarios igualmente exponencial.

En lo relativo a la captación de participantes en pruebas PT2, se considera que la clave está en el primer contacto para cerrar el acuerdo de participación entre participantes activos y el centro docente. Este contacto se puede hacer por medio de visita, por llamada telefónica o por un tercero conocido de ambos. Además de querer participar, el centro tiene que tener laboratorios con ordenadores en red disponibles para realizar las pruebas supervisadas y se debe acordar entre ambas partes las fechas y horas en las que se realizarán las sesiones. Naturalmente, no se han realizado sesiones de laboratorio durante periodos de vacaciones ni de exámenes de alumnos. Este último matiz no se ve reflejado en la Figura 40 debido a que incluso las titulaciones distintas de un mismo centro docente tienen fechas de examen distintas, lo cual complica el concretar la fecha idónea para realizar el primer contacto con el centro docente.

Asimismo, hay que destacar la importancia que tienen los coordinadores de centro. Su labor organizando los grupos, las aulas de laboratorios y demás aspectos es inestimable. La existencia de esta figura ayuda a minimizar el número de incidencias organizativas y reduce el tiempo invertido por los responsables de la administración de la PT2.

Capítulo 8
Evaluación multicriterio: CE versus CT

En este capítulo se realiza el estudio comparativo de los dos experimentos de calibración de ítems efectuados, CE y CT, considerando dos dimensiones o criterios: las estimaciones de los parámetros de dificultad de ítems obtenidas y los costes tanto temporales como económicos.

El estudio comparativo entre las estimaciones generadas por ambos procesos de calibración (sección 8.1) se centra en las estimaciones de dificultad D_i y b_i generadas por los procesos CE y CT respectivamente. Las submuestras que se comparan contienen las estimaciones de los 163 ítems comunes calibrados por ambas calibraciones. Luego, la comparación de estimaciones se restringe al parámetro dificultad y a los ítems comunes. El estudio comparativo de las estimaciones que se presenta en esta sección evalúa si existen o no diferencias estadísticamente significativas entre las estimaciones de dificultad generadas por un método de calibración u otro para un mismo conjunto de ítems.

A continuación, en la sección 8.2, previa normalización de los costes de los 4 procesos alternativos de recogida de datos ejecutados (8.2.1), se contrastan los costes temporales y económicos de las calibraciones realizadas, para lo cual se presentan diversos estudios. Primeramente, se determina la alternativa más eficiente para recabar datos entre las dos alternativas estudiadas en cada calibración: PE1 y PE2 por un lado (8.2.2) y PT1 y PT2 por otro (8.2.3). Posteriormente, se estiman y comparan los costes de las calibraciones CE y CT con sus respectivas alternativas más eficientes

de recopilación de datos. La sección finaliza extrapolando las estimaciones de costes de calibraciones con otros tamaños de bancos de ítems distintos al n=252 original (apartado 8.2.5).

8.1. Análisis de las estimaciones de dificultad resultantes

En esta sección se realiza un estudio comparativo para determinar si los dos procesos de calibración desarrollados generan estimaciones de dificultad de ítems semejantes.

El método seguido en el estudio ha sido el siguiente. Primeramente, se han unificado las escalas de las muestras de datos a emplear en el contraste. Seguidamente, se ha formulado la hipótesis declarando las pruebas a realizar. Finalmente, se han analizado los valores de las estimaciones puntuales y de intervalos obtenidos, concluyendo el estudio con una interpretación combinada de los mismos.

8.1.1. Transformación de la escala de los datos

Se dispone de 163 pares de valores de la forma (D_{ik}, B_{ik}) con $i \in [1, 252]$ (los índices de los ítems del banco original), $k \in [1, 163]$ (los índices correlativos para los ítems calibrados por CE y por CT) y $k \leq i$. El vector $<D_{ik}>$ representa las 163 dificultades estimadas por el proceso CE y con media poblacional μ_D. Análogamente, el vector $<B_{ik}>$ recoge otras tantas dificultades estimadas para los mismos ítems pero por el proceso CT, y su media se denota mediante μ_B. Los D_{ik} están en el intervalo [1, 12] y los B_{ik} en (-3.5, 3.5).

Para hacer comparables los valores D_{ik} y B_{ik}, primero se convierten **a una métrica común en el intervalo de 0 a 1**. Al no existir un procedimiento estándar de normalización de escalas, y puesto que escoger un procedimiento u otro puede incidir en los resultados del test estadístico a aplicar, se opta por convertir los vectores de valores paralelamente aplicando los 4 procedimientos descritos en la

sección 4.5. Específicamente, la aplicación del procedimiento *j* convierte los vectores $<D_{ik}>$ y $<B_{ik}>$ en los vectores $<D_k^j>$ y $<B_k^j>$. Los valores concretos de los 4 pares de vectores obtenidos se hallan tabulados en el anexo A6, así como los valores originales estimados por los procesos CE y CT.

8.1.2. Formulación de la hipótesis

A través del contraste de la hipótesis se quiere clarificar la equivalencia entre las estimaciones de dificultad de los ítems que generan los procesos CE y CT de calibración. Así, aún siendo procesos distintos, se espera que **en promedio las estimaciones de dificultad de ítems que generan sean iguales** ($\mu_D - \mu_B = 0$). La hipótesis nula (H_0) y la alternativa (H_A) quedan formuladas de la siguiente forma:

H_0::	$\mu_D - \mu_B = 0$. En promedio, las estimaciones de dificultad de los ítems determinadas por los procesos de calibración CE y CT son iguales.
H_A::	$\mu_D - \mu_B \neq 0$.

Para realizar el análisis inferencial se realizarán estimaciones puntuales y se calcularán intervalos de confianza.

8.1.3. Análisis estadístico

El cálculo estadístico se ha realizado con el paquete SPSS. El **test** que se ha usado para hacer la prueba de significancia estadística es el de **suma de rangos con signo de Wilcoxon**, ya que la variable *dificultad* es continua y se dispone de muestras dependientes sin distribución conocida. El nivel de significación se fija en (α=0.05). Así mismo, se aplica el **T-test** para comparar las medias de la muestra D_k y la B_k, construyéndose **intervalos de confianza** al 95% **para** el estadístico **media de las diferencias pareadas** ($\overline{D-B}$). Al tener muestras dependientes, los cálculos con ANOVA no serían apropiados. Para este caso, con el tamaño muestral 163, y en virtud del Teorema Central del Límite, se asume la normalidad en la

distribución de los datos obtenidos por la diferencias (D_k-B_k). La Tabla 37 recoge los resultados de aplicar la prueba de Wilcoxon y el T-test a los datos ya transformados por los 4 procedimientos indicados. Cada fila *j* de la tabla recoge el resultado de la aplicación de ambas pruebas a los datos transformados mediante el procedimiento P_j. La primera columna muestra los valores de *p calculados* por Wilcoxon con el nivel de significación establecido. En la segunda columna se recoge la equivalencia de un nivel de dificultad en el intervalo [1, 12] en el nuevo intervalo 0-1. Esta información es interesante debido a que la compactación generada entre los nuevos valores por cada procedimiento de transformación es distinta, y su conocimiento puede facilitar la interpretación de los resultados. Aclarar que el valor correspondiente al procedimiento P_2 se muestra en cursiva, ya que si bien el procedimiento de transformación respeta la cardinalidad, no garantiza que se conserve la proporcionalidad. En la columna $\overline{D-B}$ se recogen las medias muestrales de las diferencias pareadas, tal que en la fila *j* está la media de las diferencias de los valores pareados ($D_k^j - B_k^j$); en las dos últimas columnas se recogen los intervalos de confianza al 95% de correspondiente estimador así como su amplitud.

	p_calculado	**Amplitud (1 nivel dif. [1,12])**	$\overline{D-B}$	**IC.95 ($\overline{D-B}$)**		**Amplitud IC.95**
P_1	0.748	0.08982	-0.00035	-0.02240	0.02170	0.04410
P_2	0.022	*0.10248*	0.03145	0.00568	0.05723	0.05155
P_3	0.727	0.00110	0.00000	-0.00027	0.00027	0.00054
P_4	0.995	0.01318	-0.00065	-0.00389	0.00260	0.00649

Tabla 37.- Síntesis de los valores de Wilcoxon y del T-Test para muestras dependientes

Si se observan los valores de los *p calculados* por el test de Wilcoxon tras las transformaciones P_1, P_3 y P_4 (columna *p* de la Tabla 37), se ve que todos ellos superan el umbral del nivel de significación fijado. Puesto que ninguno tiene significancia estadística, la H_0 no se descartaría. Estos resultados concuerdan con los obtenidos por el T-test aplicados a las mismas transformaciones. Concretamente, en el estudio 3 se ve que la media muestral de $\overline{D^3-B^3}$ es ya 0; y para los estudios 1 y 4, si bien la medias muestrales $\overline{D^1-B^1}$ y $\overline{D^4-B^4}$ no son 0, sus correspondientes IICC al 95% sí que contienen a sus respectivas medias poblacionales (al valor 0). Luego, según estas dos transformaciones tampoco se aprecian diferencias estadísticamente

significativas. Para el estudio 2 es necesario ampliar el intervalo de confianza del estimador $\overline{D^2 - B^2}$ al 99% para que éste contenga al 0 (-0.00499, 0.06790). Por tanto, se concluye que **no hay evidencias estadísticamente significativas de que** los procesos de calibración **CE y CT produzcan estimaciones de dificultad distintas**, ya que la hipótesis nula no se rechaza, a la vista de los resultados de ambos tests aplicados.

En la Figura 45 se muestran los valores pareados resultantes de aplicar las transformaciones P_1, P_2, P_3 y P_4 a los vectores D_k y B_k originales. Los pares de valores se presentan en orden creciente de los valores D_k^j (con j en [1, 4]) que se han representado gráficamente por el icono • y, unido al mismo mediante una línea vertical, su correspondiente valor B_k^j pareado. Cuanto menor es la diferencia $D_k^j - B_k^j$, más corta es la línea que une los valores emparejados.

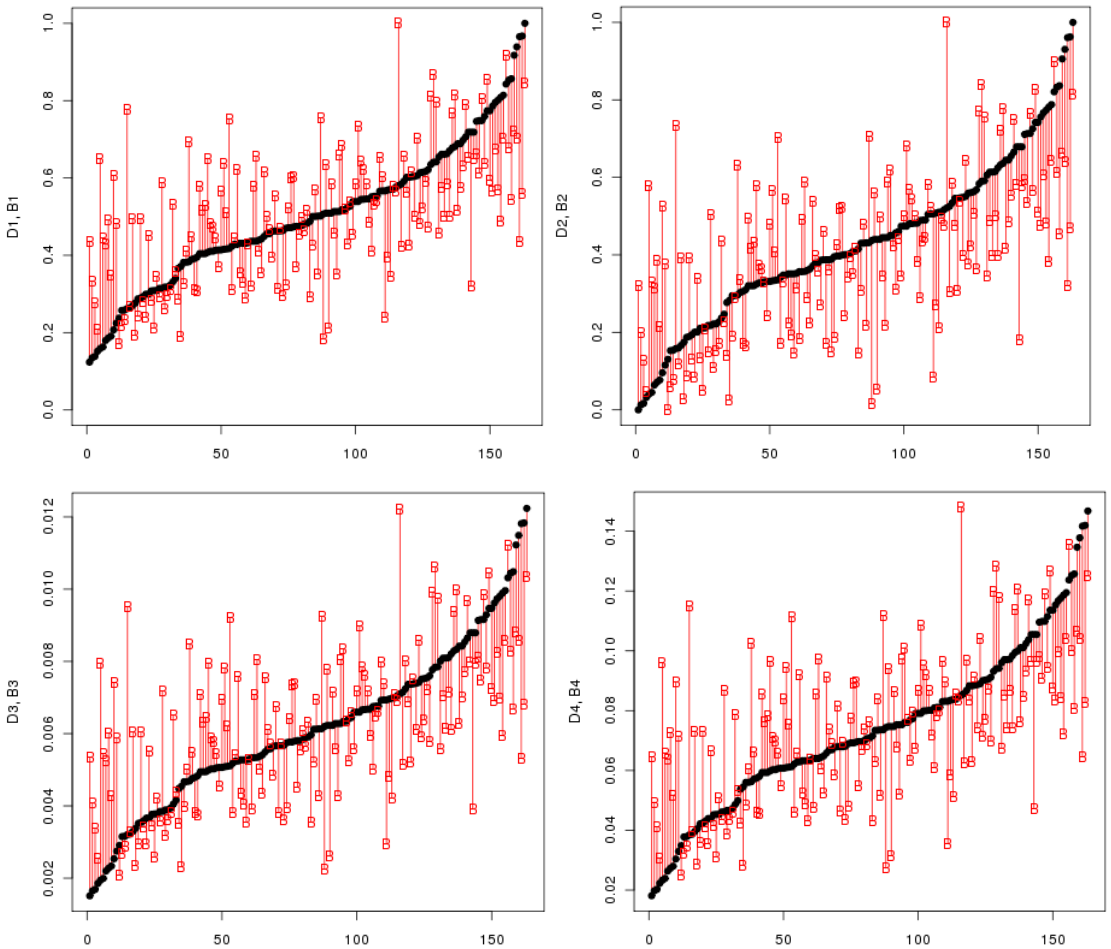


Figura 45.- Valores pareados Dk y Bk tras aplicar la correspondiente normalización de escala mediante los procedimientos P_1, P_2, P_3 y P_4

Si bien entre las calibraciones CE y CT estadísticamente no hay diferencias significativas, en los 4 gráficos se puede apreciar que las diferencias entre los 163 pares de valores mostrados existen. Sin embargo, la similitud entre las cuatro curvas y sus respectivos valores pareados es clara. Conviene precisar que la concentración de valores en las dos gráficas inferiores son características propias de los procedimientos P_3 y P_4.

De los resultados de las pruebas efectuadas (Tabla 37), también se puede estudiar la **magnitud del efecto** (*effect size*) de las estimaciones generadas por ambos procesos. Al ser cero el valor de la magnitud en el tercer estudio, no hay diferencias entre las estimaciones que generan uno y otro proceso de calibración, lo que concuerda con lo inferido hasta el momento. Para los estudios 1 y 4, como es negativa, significa que el procedimiento CT considera los ítems más difíciles en comparación con las estimaciones obtenidas por CE. Sin embargo, esas magnitudes son mínimas: la del primer estudio es 1/256 de un nivel de dificultad en la escala [1,12], mientras que la magnitud del cuarto estudio es 1/20. Con estos valores prácticamente nulos de la **magnitud del efecto** podemos inferir que **las diferencias entre las estimaciones de dificultad de CE y CT son inexistentes**.

8.2. Análisis de costes temporales y económicos

En esta sección, se muestran agrupados los costes globales de las calibraciones CE y CT, y se comentan diversos contrastes alternativos.

En la Figura 46 se muestran los costes temporales globales de las calibraciones CE y CT realizadas, habiendo requerido la calibración tradicional con expertos un total de 1220h frente a las 2933h de la estadística. Al observar los desgloses por fases de desarrollo e hitos considerados, se aprecia que los costes considerados en las segundas fases son similares; por lo que necesariamente las diferencias tan elevadas están en el desarrollo de las primeras fases. Específicamente, el hito con diferencia más acusada entre ambas calibraciones es *t. pasivos* (con 112h invertidas por expertos, frente a

las 1516h de los sujetos anónimos), seguido del apartado *implementación* donde los valores alcanzados no han sido tan elevados a pesar de que los de la calibración estadística casi duplican los de la calibración basada en expertos (396h vs 694h). El resto de apartados, al igual que los considerados en la fase posterior, no siendo iguales, sí que son parejos.

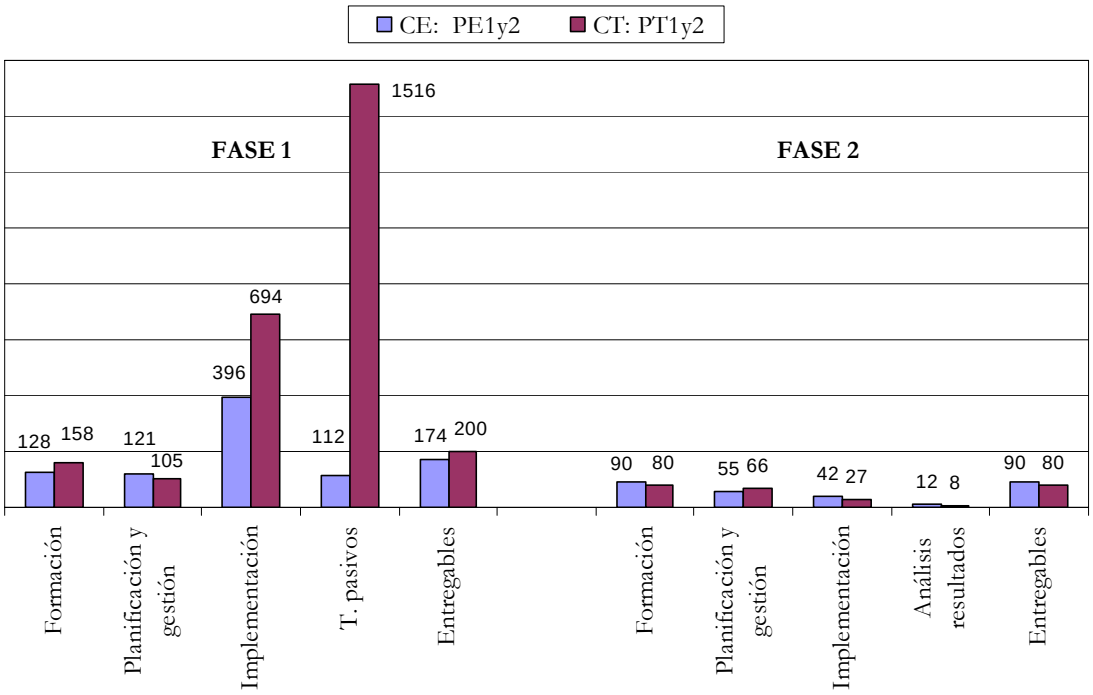


Figura 46.- Costes temporales de las calibraciones realizadas desglosadas por fases y apartados considerados

Sin embargo, es posible realizar un análisis más riguroso. Anteriormente ya se ha argumentado que para realizar una calibración eficiente desde el punto de vista beneficio-coste, se necesitan al menos 7 valoraciones de expertos para la calibración vía expertos y al menos 500 valoraciones de sujetos anónimos para la calibración 3PL-TRI. En este estudio, la calibración vía expertos se ha realizado aunando las valoraciones de las pruebas de campo PE1 y de las PE2, alcanzando, en términos medios y una vez depurada la muestra, 15 valoraciones de experto por ítem (más del doble del mínimo necesario). En el caso de la calibración estadística, ésta se ha realizado también con las valoraciones conjuntas recogidas entre las pruebas de campo PT1 y PT2, que en promedio han sido 540 valoraciones por ítem. En consecuencia, es necesario *normalizar los costes* de las estimaciones de las fases de recogida de datos a 7 valoraciones de experto y a 500 de sujetos anónimos, para hacer

comparables los costes de ambas calibraciones. Dichos procesos de normalización se realizan y argumentan en la sección 8.2.1. A partir de los valores normalizados se comparan y determinan las mejores alternativas para recabar datos a través de las pruebas de campo estudiadas tanto para la calibración CE como para la CT (secciones 8.2.2 y 8.2.3, respectivamente). Seguidamente, y a partir de las pruebas PE y PT más eficientes, en la sección 8.2.4 se presentan estimaciones de costes de las correspondientes calibraciones. En la sección 8.2.5, el contraste de costes se lleva un paso más allá, al extrapolar la estimación de costes de la opción más eficiente de la calibración con expertos y de la estadística con bancos de ítems con distintos tamaños.

8.2.1. Normalización de costes

Para contrastar los costes de las cuatro pruebas de campo ejecutadas, es necesario *normalizar* los resultados ya que no se han realizado sobre las mismas magnitudes. Concretamente, los costes de las pruebas PE se normalizan a obtener 7 valoraciones de experto por ítem y los de las pruebas PT a obtener 500 valoraciones de sujetos anónimos. En consecuencia, se normalizan los valores de los costes temporales y económicos de las pruebas PE1, PT1 y PT2.

La normalización de los valores se ha calculado *proporcionalmente* por apartados considerados y a partir de los costes medidos para obtener 10 valoraciones mediante la PE1, 162 mediante la PT1 y 378 mediante la PT2, salvo para el apartado tiempo en formación, que se ha realizado una única vez y su coste es invariable con respecto al número de valoraciones recabadas.

8.2.1.1. Valores normalizados de las pruebas PE

En la Tabla 38 se presentan los costes de la recogida de datos con expertos. En la segunda columna se presentan los costes de la PE1 tal y como aparecen en la sección 6.1.8. En la tercera columna aparecen estos valores *normalizados* para poder ser contrastados con los valores de la PE2 (última columna). Los costes económicos se presentan parametrizados, en lugar de en euros, para que sean

directamente calculables a partir de las tarifas vigentes consideradas en cada momento.

TIEMPO INVERTIDO (h)	**PE1**	**NPE1**	**PE2**
Formación	128	128	128
Planificación y gestión	108	105	95
T. pasivos	77	50	50
Implementación	302	266	248
Entregables	164	155	155
TOTAL	**779**	**703**	**675**
TELÉFONO & EMAILS	**PE1**	**NPE1**	**PE2**
N. llamadas a c. a consultados	83	53	94
T. llamadas c. consultados (h)	6	3.8	7.9
N. emails	32	20	51
DESPLAZAMIENTOS	**PE1**	**NPE1**	**PE2**
N. viajes a centros acordados	53	32	-
T. viajando a c. acordados (h)	37.6	22.6	-
Km.s desplazados a c. acordados	1576	945.6	-
CUESTIONARIOS PAPEL	**PE1**	**NPE1**	**PE2**
N. copias x cuestionario	91*33	54*33	80*33
N. franqueos postales	-	-	2*51
EQUIP. INFORMÁTICO	**PE1**	**NPE1**	**PE2**
Ordenador y sw de ofimática	1	1	1

Tabla 38.- Costes normalizados en la fase de recogida de datos con expertos

Para los estudios presentados en esta memoria las tarifas empleadas han sido las recogidas en la Tabla 39.

Concepto	**Coste**
ADSL (Telefónica)	30€/mes
Fotocopias (10 copias<)	0.042€/copia
Franqueo postal	0.32€/sello
Tarifa por km desplazado (UPV/EHU)	0.29€/km
Ordenador y Sw ofimática	400€/u.
Servidor web & licencia Internet Information Server	1000€/u.

Tabla 39.- Tarifas vigentes a fecha 4/12/2009

8.2.1.2. Valores normalizados de las pruebas PT

En la Tabla 40 se presentan los costes temporales normalizados de la recogida de datos con sujetos anónimos para la calibración CT a través de las pruebas PT. La extrapolación de los valores se ha calculado proporcionalmente a partir de los costes medidos de sus

respectivas valoraciones (explicadas en las secciones 7.1.8 y 7.2.6), exceptuando el apartado *tiempo en formación*, cuyo coste (158h) no varía, como se ha comentado anteriormente; y el apartado *entregables*, que se ha considerado de igual coste (200h) al de los entregables asociados a las 540 valoraciones reales obtenidas. Para estimar los valores de las dos primeras columnas se procede de forma análoga a la normalización de los costes de la prueba PE1 (véase la sección previa), pero en esta ocasión extrapolando los costes de las valoraciones obtenidas en una prueba concreta hasta alcanzar las 500 valoraciones fijadas.

TIEMPO INVERTIDO (h)	**NPT1**	**NPT2**	**PT1y2**
Formación	158	158	158
Planificación y gestión	114.5	116.2	104.8
T. pasivos	1903.6	991.8	1515.6
Implementación	530.2	718	693.6
Entregables	200	200	200
TOTAL	2906.3	2184	2672
TELÉFONO & EMAILS	**NPT1**	**NPT2**	**PT1y2**
N. llamadas	873.8	92.5	354
T. llamadas	58.26	13.9	29.4
N. emails	3945	-	1282
T. emails	125.1	-	40.7
DESPLAZAMIENTOS	**NPT1**	**NPT2**	**PT1y2**
Km.s desplazados a c. acordados	-	2055.8	1555.8
T. viajando a c. acordados	-	72.9	55.2
EQUIP. INFORMÁTICO	**NPT1**	**NPT2**	**PT1y2**
Ordenador y sw de ofimática	1	1	1
Servidor web y licencia Internet Information Server	1	1	1

Tabla 40.- Costes normalizados en la fase de recogida de datos con sujetos anónimos

8.2.2. Comparación de costes: PE1 vs PE2

La comparación entre los costes temporales y económicos de pruebas de PE se realiza empleando los valores normalizados presentados en la Tabla 38. Estos valores representan las estimaciones de costes para recabar 7 valoraciones de expertos.

Atendiendo al factor **tiempo invertido** (Figura 47), **la prueba PE1 es más costosa que la PE2**. Concretamente, el realizar la PE2 supone un 7% de ahorro frente a la PE1, cuando se consideran los apartados *formación*, *planificación y gestión*, *tiempo* invertido por participantes *pasivos*, *implementación*, elaboración de *entregables* y el *tiempo viajando* a los centros con los que ha habido acuerdo (específicamente 675h de la PE2 frente a las 726h de la PE1). No obstante, el recopilar datos a través de la PE2 supone ahorrar un 13% de tiempo, si únicamente consideramos los apartados con tiempos variables (*planificación y gestión*, *implementación* y *tiempo viajando*). Estos resultados confirman el presentimiento que se tenía sobre el esfuerzo superior realizado en la PE1 al ejercer un mayor control sobre el desarrollo del trabajo.

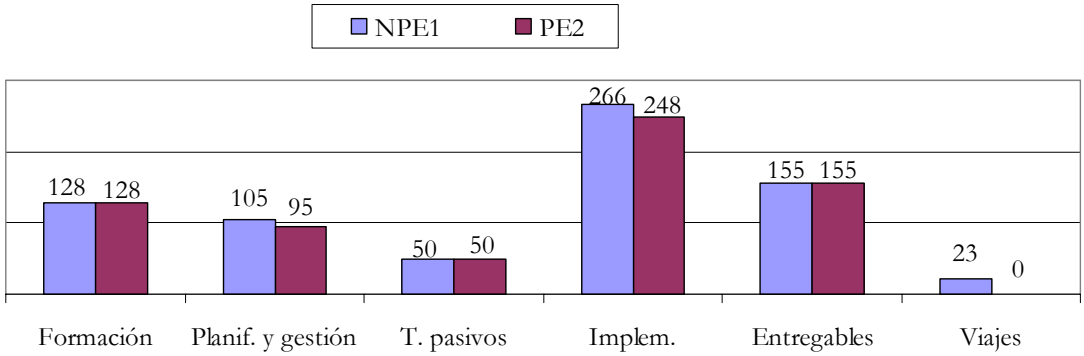


Figura 47.- Contraste de costes temporales variables en la recogida de datos con expertos

Nótese que el tiempo invertido en las llamadas telefónicas registradas en la Tabla 38 está contabilizado entre los apartados de *gestión y planificación* y de *implementación*.

Desde el punto de vista de **costes económicos,** la Tabla 38 recoge las cuantías desglosadas por apartados considerados. Ambas pruebas han precisado de un ordenador y software de ofimática. Por lo tanto, los aspectos diferenciadores desde el punto de vista de costes económicos son el *número de copias*, el *franqueo postal* y el *coste de los desplazamientos*. Si bien al ser superior la tasa de abandono por parte de los expertos en la PE2, el coste de adicional de las copias de los cuestionarios y del franqueo postal de la PE2 es inferior al coste de los desplazamientos de la PE1. Por lo que desarrollar una réplica de PE2 es más económico que conducir una PE1. Concretamente, el coste de la PE2 es un 67.2% más económica que la PE1 (153.72 euros vs 468.08 euros, según las tarifas de Tabla 39).

Luego, **para recabar 7 valoraciones de expertos, los costes económicos y temporales estimados corroboran empíricamente que la PE2 es una réplica mejorada de la PE1, siendo la PE2 más eficiente que la PE1.**

8.2.3. Comparación de costes: PT1 vs PT2

En este apartado se contrastan las estimaciones de costes para lograr 500 valoraciones de sujetos anónimos con únicamente pruebas de campo PT1 y, por otro lado, únicamente con administraciones PT2 (columnas NPT1 y NPT2 en la Tabla 40, sección 8.2.1.2).

Atendiendo al factor **tiempo invertido** (Figura 48), el volumen de horas empleadas en la PT2 es un 22.3% menor que las de la PT1 cuando se consideran los tiempos de los apartados *formación, planificación y gestión*, *participantes pasivos*, *implementación*, *entregables* y *desplazamientos* (Tabla 40). La suma total de los tiempos estimados en los 6 apartados considerados asciende a 2906h en la PT1 y a 2257h en la PT2. Nuevamente hay que indicar que el tiempo invertido en llamadas telefónicas está distribuido entre los apartados de *planificación y gestión* y el de *implementación*; en concreto, en el primer apartado se han incluido los tiempos de las llamadas de captación y organización, y en el siguiente las realizadas para validar sesiones. Luego, según estas estimaciones, **para obtener al menos 500 valoraciones de ítems la opción más eficiente es la PT2**, proceso que requiere **3096 test completados** si la tasa de descarte de sesiones acabadas sigue siendo del 3.2%.

Sin embargo, al observar dichos costes gráficamente (Figura 48), hay que reflexionar sobre si debe o no considerarse el tiempo invertido por los participantes pasivos.

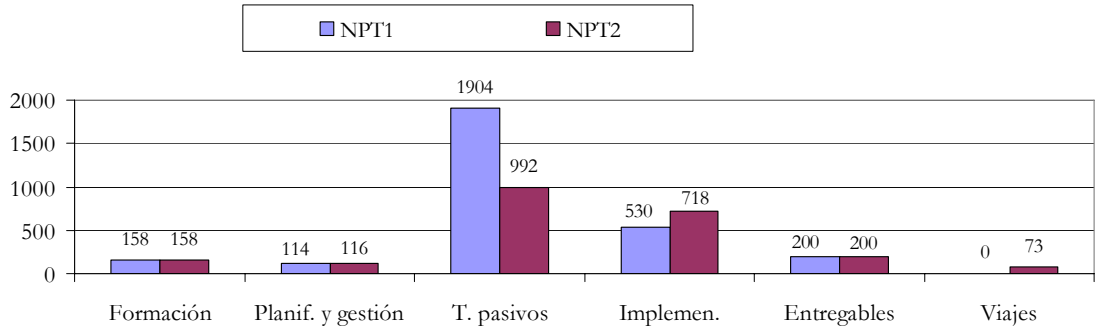


Figura 48.-Contraste de costes temporales estimados en la recogida de datos con sujetos anónimos.

En la Figura 49 se presentan las mismas estimaciones temporales desglosadas por tipo de participantes y en total. De esta forma, **si no se tuvieran en cuenta los costes asociados a los participantes pasivos**, la forma **más provechosa** de obtener las 500 valoraciones sería a través de **la prueba PT1** (1002h frente a 1265h), siendo este coste casi un 20% menor que la PT2. Sin embargo, el mayor escollo que presenta esta alternativa es que, si la tasa de sesiones (completas) rechazadas sigue siendo del 40.37%, para realizar el estudio se necesitaría que **5031 participantes** completasen el test, lo que hace que esta alternativa sea complicada de llevar a cabo.

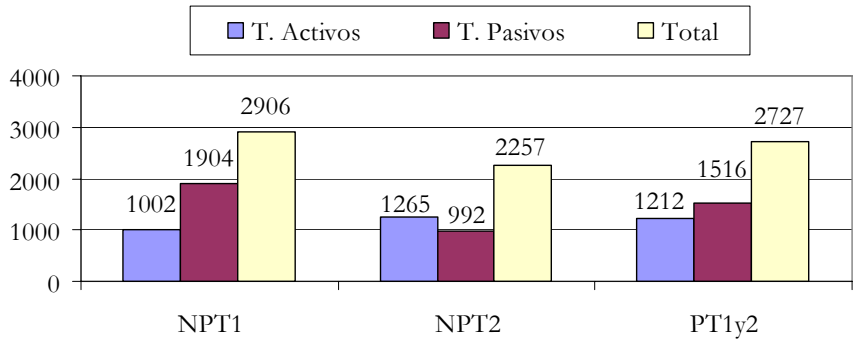


Figura 49.- Desglose de estimaciones de tiempo por tipo de participantes para obtener al menos 500 valoraciones de sujetos anónimos

Desde el punto de vista de **costes económicos**, el único apartado económico diferenciador sería el de los desplazamientos, ya que los costes telefónicos y de correos electrónicos estarían incluidos en ambas pruebas, al igual que el equipamiento informático necesario, al ser para ambas un ordenador personal con software de ofimática y un servidor web con las licencias pertinentes (véase la sección 7.1.8 o bien la 7.2.6). Por tanto, y puesto que **la PT1** no tiene desplazamientos asociados, esta **es la alternativa más económica** para recabar 500 valoraciones con sujetos anónimos.

Una vez identificados por cada tipo de calibración la variante de recogida de datos viable y más eficiente, se contrastarán los costes de realizar sendas calibraciones únicamente con valoraciones obtenidas a través de dichas pruebas de campo.

8.2.4. Comparación de costes: CE vs CT

En la Tabla 41 se recogen las estimaciones de realizar las fases consecutivas de "recogidas de datos" y de "análisis de datos y

calibración de ítems" de la calibración CE y de la CT, empleando únicamente las valoraciones obtenidas con las mejores alternativas estudiadas en las dos secciones anteriores. De este modo, la columna CE-PE2 muestra los costes de realizar una calibración con expertos con 7 valoraciones de ítems obtenidas de pruebas PE2 y la columna CT-NPT2 una calibración 3PL-TRI con 500 valoraciones de pruebas PT2. Los valores de esta última columna están normalizados ya que, en promedio, únicamente se han obtenido 378 valoraciones (véase la sección 8.2.3).

	Apartados	**CE-PE2**	**CT-NPT2**
Fase 1: **Recogida de Datos**	Formación	128	158
	Planificación y gestión	95	116.2
	T. pasivos	50	991.8
	Implementación	248	718
	Entregables	155	200
	Subtotal(h)	***676***	***2184***
Fase 2: **Análisis y calibración**	Formación	90	80
	Planificación y gestión	55	66
	Análisis resultados	12	8
	Implementación	42	27
	Entregables	90	80
	Subtotal(h)	***289***	***261***
	Coste temporal(h)	**965**	**2445**

Tabla 41.- Síntesis de horas estimadas para realizar la CE con pruebas PE2 y la CT con pruebas PT2

Si se observan los tiempos globales, la calibración estadística es un 253% más costosa que la de expertos. Al observar los apartados considerados con más detalle, la diferencia principal está en el tiempo invertido por los sujetos pasivos. En la Figura 50 se muestran los costes (normalizados) de todas las variantes estudiadas en esta memoria. En la figura queda patente que los tiempos de los sujetos activos de todas las variantes son "similares", mientras que los volúmenes atribuidos a los expertos/revisores y sujetos anónimos oscilan considerablemente de una variante de calibración a otra.

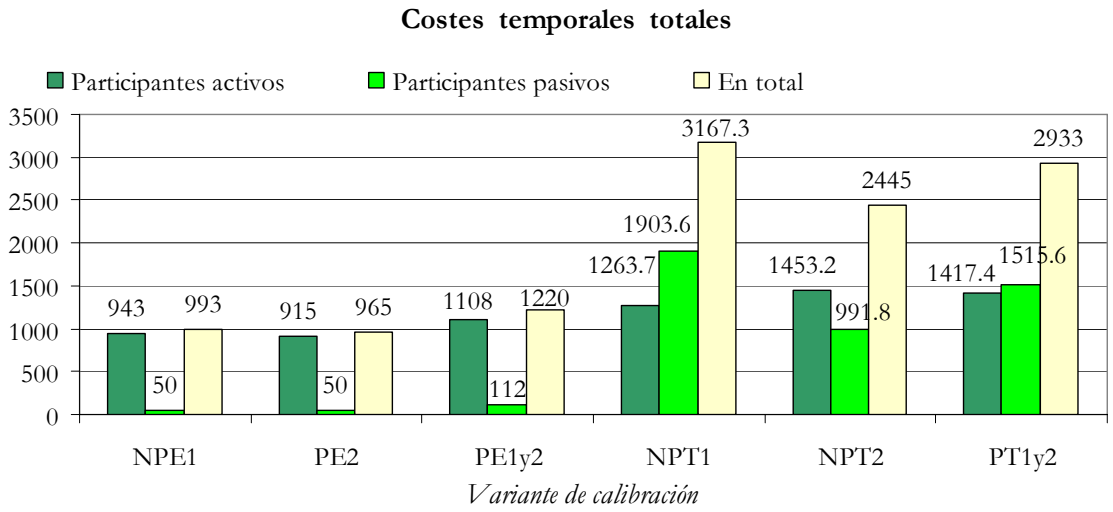


Figura 50.- Horas invertidas por participantes activos vs pasivos en las 6 variantes de calibraciones consideradas

Como balance del análisis de los costes temporales, las alternativas más interesantes de los tipos de calibraciones consideradas son la variante PE2 en la fase de recopilación de datos para la calibración con expertos; y la variante PT2, para el caso de la calibración 3PL TRI, calibraciones que se designarán de este punto en adelante por CE-PE2 y CT-NPT2, respectivamente. Bien si se incluyen todos los apartados temporales considerados, bien si únicamente se consideran los apartados atribuidos a los participantes activos, **desde el punto de vista de costes temporales, la alternativa más eficiente para calibrar un banco con 252 ítems es la calibración CE-PE2**. El volumen de horas a invertir por los participantes activos es un 37% menor en la CE-PE2 en cualquiera de las dos opciones barajadas, tanto si se consideran todos los apartados como si no se consideran los tiempos invertidos por los participantes pasivos, bien sean expertos o sujetos anónimos (Figura 50).

En cuanto a los costes **económicos** de las variantes CE-PE2 y CT-NPT2, los conceptos de gasto considerados en ambas alternativas se han listado en la Tabla 42. Suponiendo que los ciclos de vida de ambas calibraciones fueran iguales, y dejando de lado los conceptos comunes a ambas variantes, **los costes** remanentes **de la calibración con expertos son inferiores a los de la calibración 3PL**. De hecho, el coste por las copias de los cuestionarios junto con los franqueos postales son inferiores a los costes por

desplazamiento requeridos por la PT2, el servidor web y las licencias software requeridas para realizar la calibración estadística.

Concepto	CE-PE2	CT-NPT2
Tarifa plana ADSL (Llamadas telefónicas e Internet)	X	X
Desplazamientos		X
Copias de cuestionarios	X	
Franqueos postales	X	
Ordenador personal	X	X
Software de ofimática (con base de datos y hojas de cálculo)	X	X
SPSS (licencia sw.)	X	X
Servidor web y licencia Internet Information Server		X
Conexión a Internet del servidor		X
LISREL/PRELIS (licencia sw.)		X
XCALIBRE (licencia sw.)		X

Tabla 42.- Conceptos gastados

Llegados a este punto, podría resultar interesante analizar cómo varían los costes de cada una de estas dos opciones a medida que cambian los tamaños de los bancos de ítems a calibrar.

8.2.5. Extrapolación de costes a otros tamaños de bancos

En esta subsección se presentan estimaciones de costes temporales para construir calibraciones CE y CT con bancos de diversos tamaños. Siendo el tamaño del banco original a calibrar de n=252 ítems, las variaciones que se han considerado han sido: la mitad del tamaño del banco inicial y los múltiplo 2*n, 3*n y 4*n del mismo. Estos tamaños son volúmenes cercanos a los valores 125, 250, 500, 750 y 1000 ítems, y sirven para dar una perspectiva de las tendencias de los costes dependiendo del volumen de ítems del banco a calibrar.

Las estimaciones de costes se realizan bajo los supuestos de obtener 7 valoraciones de experto por ítem y 500 de sujetos anónimos por ítem. Igualmente, las estimaciones se realizan conservando las longitudes y formatos originales de los cuestionarios: 42 ítems para los cuestionarios en papel de CE y 60 para los cuestionarios electrónicos de CT. El reparto de los ítems

entre cuestionarios, determina el número de cuestionarios distintos a elaborar; esto, junto con el número de valoraciones a recopilar, y las tasas de abandono (48.1%) y descarte de expertos (4.3%), y la tasa de no validación de sesión supervisada (3.2%), todas ellas identificadas en los correspondientes procesos de calibración desarrollados, fijan el número de cuestionarios completados a recuperar y el número de sujetos pasivos a captar. Los valores concretos se pueden consultar en el anexo A6.

Con independencia de considerar o no el tiempo invertido por los participantes pasivos, el volumen de éste incide proporcionalmente en el tiempo de gestión e implementación de los participantes activos. En la Tabla 43 las filas CE(*p*) y CT(*p*) muestras las estimaciones de las horas que tendrían que invertir en total los participantes pasivos para calibrar bancos de ítems con los tamaños considerados. En el caso de los expertos irían desde 26 hasta 207 horas, mientras que en el caso de la TRI desde las 495 hasta las 3963 horas.

	0.5*n	**n**	**2*n**	**3*n**	**4*n**
CE(*a*)	715	915	1237	1559	1880
CT(*a*)	903	1143	1548	1952	2357
CE(*p*)	26	52	103	155	207
CT(*p*)	495	991	1982	2972	3963
CE(*total*)	741	967	1341	1714	2087
CT(*total*)	1399	2133	3529	4925	6320

Tabla 43.- Estimación de horas a invertir por participantes activos (a), pasivos (p) y en total para elaborar calibraciones CE y CT con tamaños de bancos de ítems varios

Para estimar los tiempos de los participantes activos, se han tenido en cuenta la existencia de costes fijos y costes variables. Los *costes fijos* son apartados cuyos valores son independientes del volumen del banco de ítems a calibrar, y por ello, sus valores son invariables en todas las estimaciones, como por ejemplo, el tiempo invertido en *formación*. Los *costes variables* son apartados cuyos valores aumentan conforme crecen los tamaños de los bancos de ítems a calibrar, ejemplo de ello son los entregables y los tiempos dedicados al análisis de los datos. Así, mientras que los costes fijos se han conservado, los valores de los costes variables se han calculado por aplicación de dos factores correctores con respecto a los costes variables del banco original (n=252). Los valores concretos aplicados a los *entregables* y a aspectos de *diseño* para los tamaños del

banco considerados son (0.75; 1; 1.25; 1.5 y 1.75); mientras que los aplicados a aspectos de *planificación-gestión* y de *implementación* son (0.5, 1, 2, 3, 4). Las filas CE(*a*) y CT(*a*) de la Tabla 43 recogen numéricamente los tiempos estimados a invertir por los participantes activos de las dos calibraciones consideradas. **En todos los tamaños de bancos considerados el tiempo a invertir en la calibración CE por los participantes activos es inferior**, lo que supondría un ahorro cercano al 20% frente a la calibración CT (primera fila en la Tabla 44). Esta diferencia se muestra gráficamente a través de las pendientes de las curvas de la Figura 51. En la misma figura mediante líneas punteadas se muestran las correspondientes tendencias polinomiales de cada una de las curvas.

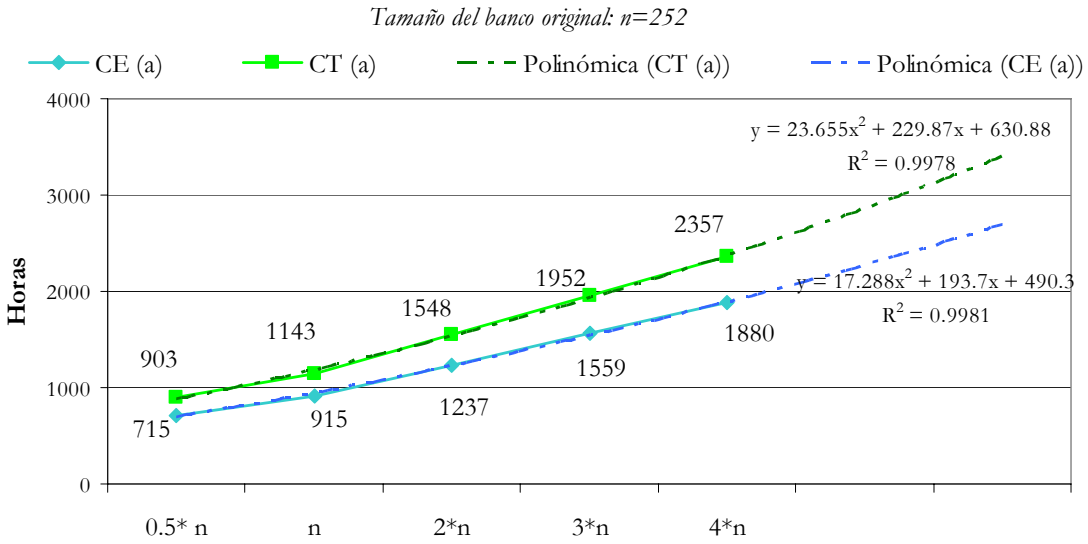


Figura 51.- Costes a invertir por participantes activos para elaborar calibraciones CE y CT con tamaños de bancos de ítems alternativos

Al considerar conjuntamente los costes de todos los participantes, la diferencia de horas entre ambos tipos calibración es más prominente, siendo siempre más eficiente la opción CE (últimas dos filas de la Tabla 43). En concreto, con un tamaño de banco de ítems 0.5*n, la calibración CE requiere un 47% menos de tiempo que lo precisaría la CT, y conforme aumenta el volumen del banco la diferencia va en aumento siempre a favor de la CE, llegando a alcanzar el 67% de ahorro. Los ahorros de tiempo que se pueden obtener por realizar una calibración con expertos frente a una 3PL para los tamaños de bancos considerados se han recogido en la Tabla 44. Concretamente, la fila *Ahorro (a)* recoge los ahorros de los tiempos por parte de los participantes activos y la fila

Ahorro (total) recoge los ahorros de tiempo por parte de los participantes involucrados, tanto los activos como los pasivos.

CE contra CT	0.5*n	n	2*n	3*n	4*n
Ahorro (a)	20.8%	19.9%	20.1%	20.1%	20.2%
Ahorro (total)	47.0%	54.7%	62.0%	65.2%	67.0%

Tabla 44.- Ahorro de tiempo por participantes activos (a) y en total

El ahorro entre los costes estimados de realizar una calibración CE frente a una CT queda patente en la Figura 52 a través de la pendiente más acusada que presenta la curva de la calibración CT. Las correspondientes tendencias polinómicas de segundo grado de cada una de las curvas se han incorporado mediante líneas punteadas.

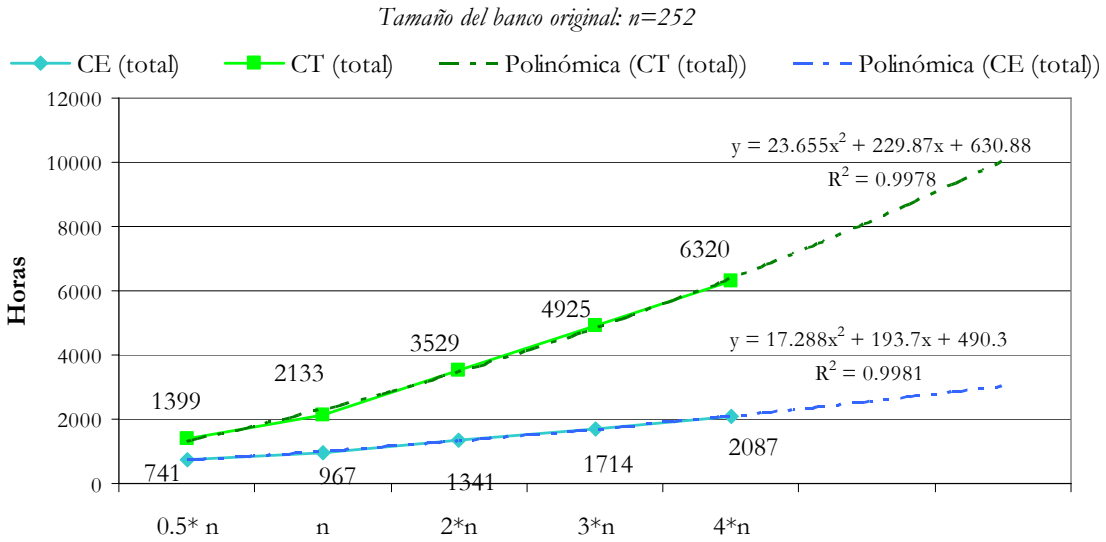


Figura 52.- Costes a invertir por participantes activos y pasivos para elaborar calibraciones CE y CT con tamaños de bancos de ítems alternativos

En el anexo A6 se pueden consultar desglosados por apartados considerados los tiempos estimados para realizar las calibraciones CE y las CT para los tamaños de bancos considerados.

Para completar la extrapolación de costes a otros tamaños de bancos, se podría añadir la estimación de los **costes económicos** para las variantes CE y CT. Los costes a confrontar son, por parte de la CE, las copias en papel de los cuestionarios y los franqueos, y por parte de la CT, los costes del servidor web, su correspondiente licencia software, los desplazamientos y las licencias de LISREL/PRELIS y XCALIBRE para el procesamiento estadístico de los datos recopilados. Si bien, los costes del servidor web y las licencias representan costes fijos para las calibraciones CT, no

sucede lo mismo con los desplazamientos: a mayor volumen de sujetos anónimos a administrar, más grupos de laboratorios que habrá que organizar y supervisar, y mayor será el número de desplazamientos que habrá que realizar. Y estos costes distan de los correspondientes a las copias en papel y el franqueo postal que precisarán las administraciones PE2. Consecuentemente, **las calibraciones CE resultan siempre más baratas que las CT**.

Capítulo 9
Propuesta de procedimiento de calibración

Con el objetivo de mejorar el desarrollo de sistemas de enseñanza online, en este capítulo se presenta una propuesta de proceso de negocio abstracta -independiente del tipo de calibración- para realizar calibraciones de *n* ítems off-line. El proceso puede instanciarse para dar lugar a una calibración basada en las valoraciones de expertos o bien para dar lugar a otra para calibrar ítems según el modelo logístico de tres parámetros de la TRI. La propuesta integral se fundamenta en la experiencia de las dos calibraciones desarrolladas y en los resultados de los contrastes presentados en esta memoria.

Los procesos de negocio se representarán gráficamente utilizando la *notación BPM* (del inglés Business Process Management notation), una **notación estándar** para el modelado de flujos de procesos de negocio y servicios web. Es una propuesta consensuada realizada en 2004 por el grupo de trabajo de BPMI (del inglés Business Process Management Initiative) grupo que representa a un gran sector de la comunidad de modelado de procesos de negocio. El estándar surge de la extracción y combinación de ideas de otras notaciones existentes (White, 2004), como los *diagramas de actividad,* los *procesos de negocio de computación distribuida de objetos de empresa (EDOC)* de UML, *clases del lenguaje de modelado de empresas (IDEF)*, el *esquema de especificación en XML de negocio electrónico (electronic business XML BPSS)*, el *Diagrama flujo de decisiones-actividades (ADF)*, *RosettaNet*, *LOVeM*, y

Event-Process Chains (EPCs) (BPMI.org, 2004). Esta notación permite representar gráficamente flujos de trabajo con distintos niveles de especificidad e incorporar roles, eventos y documentos propios de los procesos de negocio. En el anexo A7 se agrupan y se describen los elementos principales de la notación BPM a la hora de definir *Diagramas de Procesos de Negocio*. En (Arruabarrena, López-Cuadrado et al., 2010; Arruabarrena y Pérez, 2010) se han plasmado con notación BPM los procesos de las pruebas de campo PE1, PE2, PT1 y PT2 conducidas así como de las dos calibraciones desarrolladas.

El capítulo está organizado de la siguiente forma. En la primera sección se identifican las tres fases en las que se descomponen las calibraciones de ítems y que se describen en los siguientes subapartados. De este modo, primero se describen los aspectos de planificación iniciales que hay que concretar antes de proceder con una calibración (sección 9.1.1). A continuación, se proponen procesos para afrontar tanto la fase de recogida de datos (sección 9.1.2) como la de análisis y calibración de los ítems (sección 9.1.3). En la siguiente tabla se enumeran los procesos principales identificados así como los diversos agentes involucrados en un proceso de calibración de ítems:

Procesos de negocio	**Agentes involucrados**
Calibración de ítems	Supervisor general y responsable del proceso integral
Planificar prueba de campo	Coordinador y ejecutor principal
Realizar pruebas piloto Ejecutar prueba de campo Conducir prueba de campo	Otros colaboradores activos: informático, transcriptor de cuestionarios en formato papel, ayudantes para la conducción de las pruebas de campo por centros de trabajo o bien en laboratorios
Administrar cuestionario Completar cuestionario	Revisores: del material pedagógico, de los cuestionarios y de la aplicación informática administradora de cuestionarios electrónicos
Hacer análisis previos Obtener parámetros	Sujetos administrados: expertos o bien individuos anónimos
Hacer análisis posteriores	

Tabla 45.- Procesos y agentes involucrados en calibraciones de ítems

9.1. Proceso integral de calibración de ítems

Todo procedimiento de calibración de ítems se realiza en tres fases (Figura 53). Primero, se debe fijar qué tipo de calibración se quiere hacer y seguidamente se planifican las tareas a realizar, que siempre deberán incluir una segunda fase de *recogida de datos* y otra tercera de *análisis y calibración*. Las últimas fases pueden dilatarse en el tiempo hasta que haya terminado la segunda o hasta que se haya alcanzado un volumen suficiente de datos.

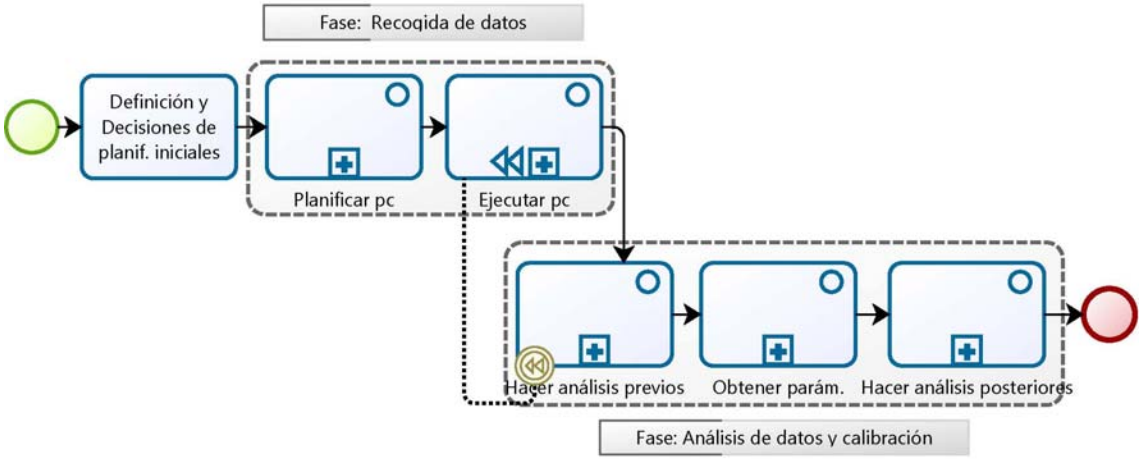


Figura 53.- Proceso de negocio "Calibración de ítems"

En la actividad **Definición y Decisiones de planificación iniciales** se identifican los objetivos del proceso de calibración, se estudia el entorno donde se va a desarrollar, y se definen los entregables que se van a producir y las herramientas a utilizar. En esta actividad también se debe incluir la formación de los sujetos. Se prosigue con el desarrollo de la fase de **Recogida de datos**, a través de la cual se recabará un volumen de datos establecido en la fase anterior. Finalmente, en la fase de **Análisis de datos y calibración** de ítems se depura la muestra de datos y se estiman los parámetros de los ítems. A continuación se describe cada una de las etapas aquí enumeradas.

9.1.1. Definición y decisiones de planificación iniciales

En esta fase se decide qué *tipo de calibración* se va a realizar (expertos o TRI), el *rasgo de los ítems* a calibrar (discriminación, dificultad, pseudoacierto, destreza...), la escala o *rango de valores* que deben aportar los *expertos* y el *rango de los valores de los parámetros que estimará el proceso de calibración*. Además de los rasgos más frecuentes, se puede aportar otra información para clasificar los ítems, como el área y subárea al que pertenece el ítem, la competencia que mide dentro de esa subárea, como por ejemplo el ritmo asociado a ítems musicales, la capacidad espacial en ítems de dibujo, etc.

Del mismo modo, se define también si se desean obtener algunos indicadores asociados al proceso de calibración (como recursos utilizados, tiempo de ejecución, etc.). Es recomendable planificar y documentar el progreso de cada uno de los procesos ejecutados, con vistas a poder acometer mejoras posteriormente o bien hacer balance del proceso íntegro.

9.1.2. Recogida de datos

Esta fase se desdobla en dos procesos. En primer lugar se *planifican* las tareas a realizar y luego se *ejecutan* (Figura 53). Estas tareas se comentan con más detalle en los siguientes párrafos.

La **planificación de la prueba de campo** se realiza por dos actores principales: el *desarrollador principal* y los *revisores*. Se define cómo alcanzar las metas propuestas estableciendo la estrategia a seguir de entre las comentadas en el Capítulo 3. Así, se debe *analizar y establecer la logística de la prueba de campo* (Figura 54) *identificar a los sujetos que van a participar* en el experimento y sus atribuciones y *diseñar los cuestionarios* a utilizar. Para asegurar la idoneidad de la extensión de los cuestionarios y localizar posibles deficiencias y carencias en los mismos es conveniente realizar *pruebas piloto*.

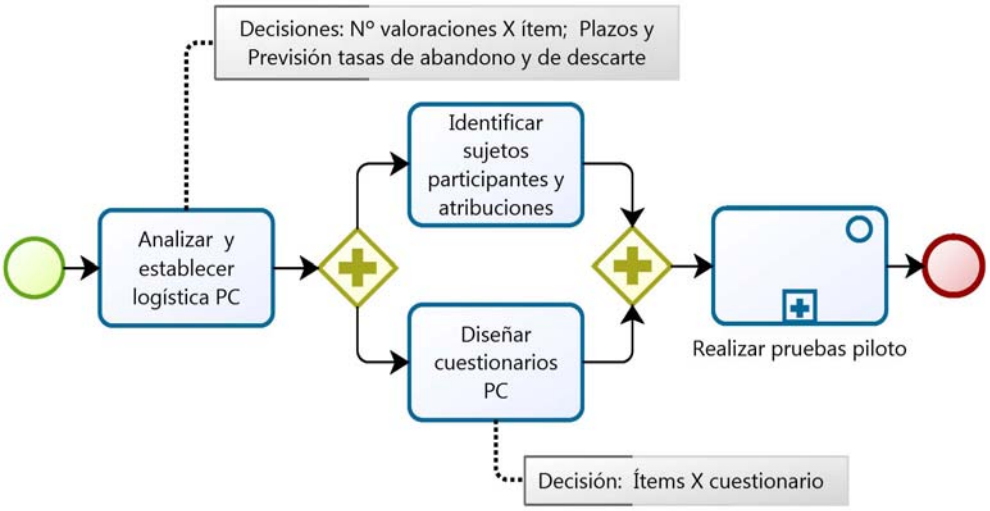


Figura 54.- Detalle del proceso de negocio "Planificar Prueba de Campo"

La actividad **ejecutar prueba de campo** consiste en *conducir* el experimento según lo planificado tal y como muestra la Figura 55. El desarrollador principal será el encargado de *gestionar las incidencias*. Si se producen imprevistos, estos se deben analizar y darles solución. También será el responsable de la recopilación y *transcripción de las respuestas* para su posterior tratamiento. La tarea de conducción se comenta más adelante con más detalle.

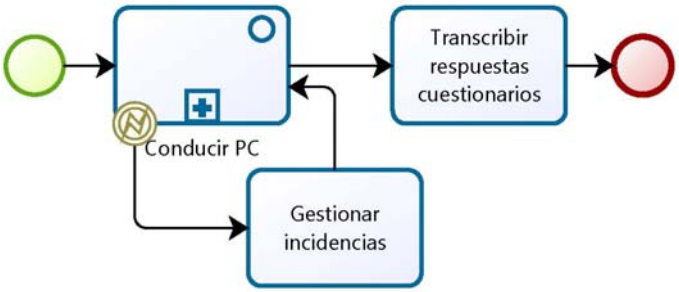


Figura 55.- Detalle del proceso de negocio "Ejecutar prueba de campo"

9.1.2.1. Análisis y establecimiento de la logística de la prueba de campo

En esta fase se deben *escoger* las *herramientas y técnicas* a utilizar para el desarrollo de la prueba de entre las definidas en el Capítulo 3, *decidir* el *número de valoraciones por ítem* que se precisan recabar, *estimar* las *tasas de abandono y de descarte* de los sujetos participantes, *establecer* el *número de revisores* que participarán en las pruebas piloto, *concretar* los distintos *plazos* a cumplir, *definir criterios para la toma de decisiones* y *acreditar los ítems de partida.*

Por ejemplo, para la calibración CE se propone recopilar la información *entrevistando* a expertos (Dix, Finlay et al., 1998; Tessmer, 1993) mediante *encuestas* plasmadas en *cuestionarios de papel*, ya que se desea evitar que la alfabetización informática de los expertos incida negativamente en la participación. Si la participación de los expertos no es remunerada, se recomienda que los cuestionarios se diseñen para poder ser respondidos en unos 45 □□ Figura 41.- Valoraciones por ítem recogidas en *la PT2* □ *ones* de experto por ítem con 48,1% como tasa de abandono y 4,3% de descarte. Para asegurarse las 7 valoraciones por ítem será preciso recopilar 8 valoraciones por ítem e involucrar a 15 expertos. En cuanto a los plazos, se propone conceder 4 semanas para que un experto complete un cuestionario y 4 meses para completar la ejecución de la recogida de datos.

Para la CT, sin embargo, se recomienda recoger datos usando *cuestionarios electrónicos* con sujetos anónimos voluntarios *en sesiones* de *laboratorio supervisadas*, puesto que el elevado número de participantes previsto no hará viable la opción de utilizar el formato de papel. Se propone que los cuestionarios tengan una longitud para que los sujetos puedan contestarlos en 20-25 minutos y su estancia en el laboratorio no se prolongue más de 40 minutos. Habrá que incluir también el tiempo dedicado a las instrucciones de completado de los cuestionarios. Si la muestra a recoger tiene que contar con al menos 500 valoraciones por ítem una vez depurados los datos, y considerando una tasa de abandono del 0% y una tasa de descarte en sesiones supervisadas acabadas del 3.2% será preciso obtener al menos 517 administraciones de cada ítem.

9.1.2.2. Identificación de sujetos participantes

Esta actividad establece qué sujetos se encargan de llevar el proceso adelante concretando cuáles son sus competencias, que normalmente irán ligadas con el resto de actividades a desarrollar. Generalmente se distinguen dos tipos de sujetos: *activos* y *pasivos*. Los sujetos pasivos normalmente son aquellos que aportan la información a recabar, mientras que los sujetos activos son los encargados de llevar adelante todos los procesos necesarios para

realizar la prueba de campo, incluyendo la captación de los sujetos pasivos.

La *calidad de los sujetos* es un tema crítico a la hora de desarrollar la prueba de campo. Si no se identifica adecuadamente al sujeto o si su participación no es correcta, el valor de sus aportaciones se verá mermada, lo que irá en detrimento de la calidad del resultado de la calibración generada.

La *captación de los participantes* se puede realizar a través de centros de trabajo, de estudios o de cualquier método que reúna gente a través de sus aficiones o responsabilidades. Es conveniente confeccionar una lista de sujetos a contactar o elementos que los reúnan (centros de trabajo o de estudios, listas de distribución, grupos de trabajo, etc.) y establecer un sistema metódico de contacto con cada uno de ellos y control de su participación.

Por ejemplo, en la PT2, los sujetos administrados se captaron a través de centros concretos (de trabajo, culturales, sociedades, instituciones, empresas o similares) que podían estar interesados en el estudio intentando que conformaran una muestra representativa. Necesariamente dichos centros debían estar dotados de salas con equipos informáticos conectados a Internet para poder acceder a la aplicación de administración de los subtests que en este contexto se denominan *laboratorios*. Por ello, entre los sujetos activos, además del desarrollador del estudio, participaron el *coordinador del centro* y los *colaboradores*. Mientras que el primero se encargó de organizar los grupos dentro de un centro y coordinarse con el desarrollador y los colaboradores para implementar la supervisión de las sesiones de laboratorio, los segundos se encargaron de identificar a los sujetos de la prueba de campo adecuadamente y de verificar las condiciones de administración de los ítems.

9.1.2.3. Diseño de cuestionarios

El **diseño de cuestionarios** consiste en fraccionar el banco de ítems en varios grupos, *decidir el número de ítems* a incluir *por cuestionario*, *qué datos* sobre los ítems hay que *recoger* y, si es necesario, la introducción de *ítems de anclaje* en cada grupo. Como se ha comentado en el Capítulo 5, los ítems de anclaje son un subconjunto del banco de ítems comunes a todos los cuestionarios y sirven para

correlacionar valoraciones de distintos participantes en diferentes cuestionarios y para establecer relaciones entre las medidas realizadas a partir de los distintos grupos de ítems. Puede ser de gran utilidad para esta tarea *determinar el tiempo medio* que un sujeto necesita para aportar la información solicitada *por cada ítem* así como conocer un tiempo máximo a dedicar *por cada cuestionario*.

Del mismo modo que ocurría a la hora de seleccionar los administrados, la confección de los cuestionarios es crucial para asegurar el éxito de la prueba de campo. Los cuestionarios se administrarán a un volumen considerable de personas, por lo que es preciso que únicamente se les solicite la información relevante para realizar el proceso de calibración; y debe asegurarse su corrección desde el punto de vista tanto de contenido como de formato. Se sugiere que la estructura de los cuestionarios esté formada por: una *introducción* que incluya directrices de compleción y ejemplos ilustrativos, un apartado de *recogida de datos personales* para fines estadísticos, el *subconjunto de ítems a valorar* y un último apartado para aportaciones *propias*. Las instrucciones deben aclarar el tipo de valoración que se solicita al experto (optimista, neutra o pesimista). En el anexo A2 se puede consultar, a modo de ejemplo, uno de los cuestionarios administrados a los expertos de la CE y en el anexo A4 uno de los 6 subtests de la PT2.

9.1.2.4. Pruebas Piloto con revisores

El objetivo de esta tarea de la fase de recogida de datos es asegurar la idoneidad de la extensión de los cuestionarios y localizar posibles deficiencias y carencias en los mismos. La Figura 56 muestra la lista de tareas propuestas para la actividad de pruebas piloto.

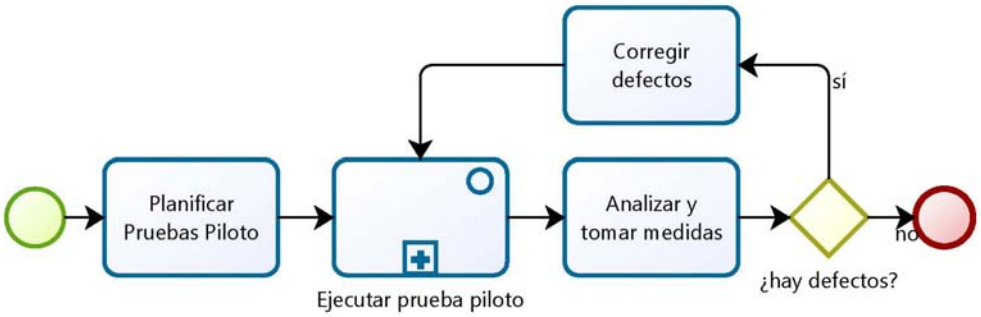


Figura 56.- Detalle del proceso de negocio "Realizar pruebas piloto"

Se propone comenzar con un primer paso de *planificación*. Seguidamente otro de *ejecución,* donde se administran los cuestionarios a revisores, quienes detectarán posibles problemas. Como se comenta en el Capítulo 5, se recomienda que participen entre 3 y 5 revisores. Cada revisor aporta una lista de defectos detectados, que se *analizan* y para los que se tomarán las *medidas de corrección* oportunas. La ejecución con revisores se repetirá en tanto en cuanto se sigan detectando defectos o el desarrollador principal considere que los mismos no tienen suficiente entidad o no hay suficiente consenso para tenerlos en consideración.

Por ejemplo, en las CE se utilizaron 5 revisores para comprobar errores de estilo en la redacción de los ítems y ciertas correcciones así como comprobar la adecuación de los cuestionarios al tiempo fijado. En la CT, además, los revisores verificaron tanto el correcto funcionamiento de la herramienta de administración de los cuestionarios como la adecuación del cuestionario de forma electrónica para todo tipo de usuarios, por lo que el número de revisores fue más numeroso.

9.1.2.5. Conducir la prueba de campo

La actividad comienza *captando los sujetos* pasivos a los cuales se les va a *administrar un cuestionario* (Figura 57). Cuando finalice el plazo preestablecido para la conducción de la prueba de campo (evento *plazo expirado*), si se ha recabado el número objetivo de cuestionarios la conducción se dará por finalizada. En caso contrario, habrá que establecer nuevas administraciones a otros sujetos y, si fuera necesario, extender el plazo de la prueba de campo. Si no hay suficientes cuestionarios ni alternativa de prorrogar la conducción de la prueba, ésta habrá fracasado (evento *Extender plazo*).

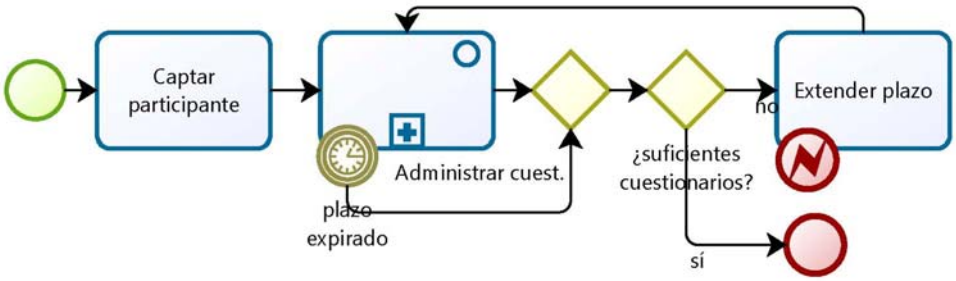


Figura 57.- Detalle del proceso de negocio "Conducir PC"

La **administración de un cuestionario** se compone de una serie de actividades que se pueden observar desde dos puntos de vista representados en la Figura 58 y la Figura 59. La primera representa las tareas que el desarrollador llevará a cabo y la segunda las tareas de los sujetos participantes pasivos. La comunicación entre ambos procesos se realiza a través de eventos con la misma designación. Ambas figuras se comentan con más detenimiento en los siguientes párrafos.

Figura 58.- Detalle del proceso de negocio "Administrar cuest."

En la Figura 58 se ve que, dependiendo del método elegido para la administración (electrónico/manual), se utilizará un método acorde para *enviar el cuestionario* (por ejemplo: Internet+correo electrónico/ papel+correo ordinario). Asociado a cada administración, se definirá un plazo correspondiente para obtener el cuestionario (evento *expiración*), que puede utilizarse para enviar un aviso de que el tiempo se está acabando, si los plazos son largos. Si *expira el plazo*, se debe *gestionar el abandono del participante* y, por tanto, no considerar su participación. Si no, se contará con el *cuestionario rellenado* y podrá formar parte de los datos *recogidos de cuestionario*.

Desde el punto de vista del sujeto participante, tras *recibir el cuestionario* (Figura 59), éste puede empezar a *completarlo*. En cuanto lo complete, *enviará el cuestionario rellenado*; si no, simplemente su participación se considerará finalizada. Se puede establecer un umbral anterior al plazo de expiración para que el sujeto participante *reciba un aviso* recordatorio (evento *plazo próximo*).

A modo de ejemplo, el proceso que se ha seguido para la ejecución de la PE2 ha sido por correo ordinario y en papel. Se estableció en 28 días el plazo para la administración de un

cuestionario. El umbral anterior a la finalización del plazo ha sido 21 días después del envío del cuestionario. Otra alternativa ha sido la empleada en las pruebas PT2, para la cual se implementó una aplicación que administraba cuestionarios electrónicos.

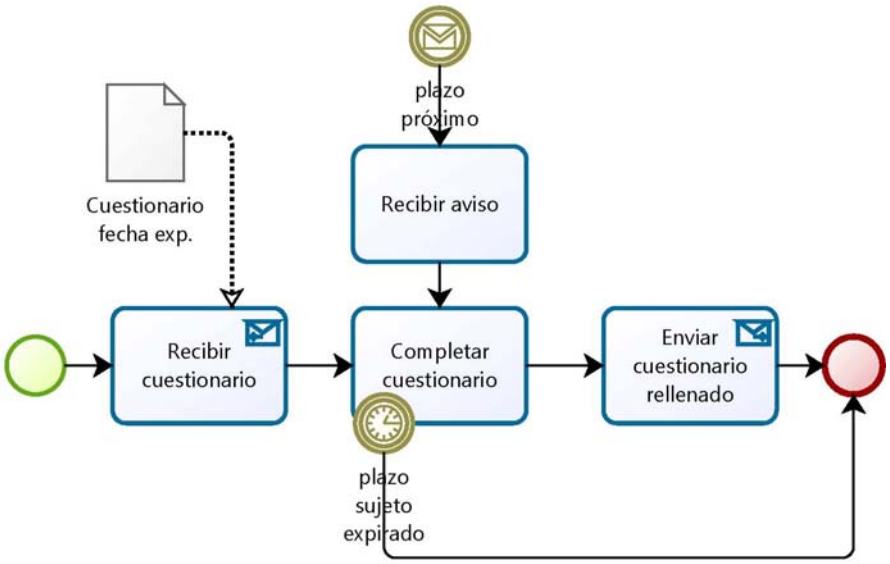


Figura 59.- Detalle del proceso de negocio "Completar Cuestionario"

9.1.3. Análisis de datos y calibración

Esta fase está compuesta de tres actividades (Figura 53). Se comienza realizando una serie de **análisis previos** para desechar aquellos datos que no superan ciertos umbrales de calidad especificados. El depurado de la muestra redundará en una mejor **estimación de los rasgos**, actividad que se hace seguidamente. Para terminar, se realizará otra serie de **análisis** que verificarán la valía de las estimaciones generadas. Las siguientes secciones describen estas actividades con más detalle.

9.1.3.1. Análisis previos

Como se ha comentado anteriormente, esta fase puede comenzar una vez que se cuenta con un cierto volumen de datos, es decir, que no es necesario esperar a que termine la fase de recogida de datos.

Los **análisis previos** permiten adelantar algunas actividades de salvaguarda de la validez experimental del proceso de calibración y apartar datos que son irrelevantes, erróneos o anómalos y que pueden malograr los resultados de la calibración de los ítems (Figura

60) . Se establecen y aplican *filtros* y *operaciones* para depurar los datos y verificar pautas de respuesta anómalas. Asimismo, en la calibración estadística y para detectar ítems incompatibles con el modelo estadístico a emplear (Renom y Doval, 1999), se efectuará un *análisis clásico de ítems* (en términos de la TCT) que descartará ítems con características extremadamente desfavorables. También se propone adelantar la comprobación del supuesto de la *unidimensionalidad* y el análisis del *funcionamiento diferencial de los ítems*. Establecidos los filtros, las operaciones y los análisis a realizar, se ordenan y se concretan los que tienen que plasmarse de forma combinada y/o cíclica.

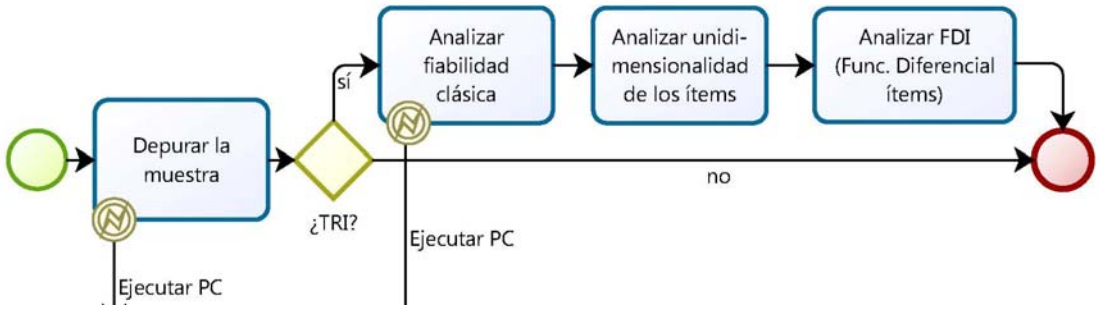


Figura 60.- Detalle del proceso de negocio "Hacer análisis previos"

La depuración de los datos y los análisis pueden reducir el volumen de la muestra por debajo del umbral de datos establecido para realizar la calibración, lo cual puede desencadenar una tarea de *compensación* (evento de la Figura 53 y Figura 60). Generalmente la compensación supone dilatar la tarea de recogida de datos hasta alcanzar el umbral de valoraciones fijado. Es más, es hasta recomendable iniciar la depuración de la muestra antes de finalizar la recogida de datos, ya que en esta fase se pueden desechar datos recogidos y una consecuencia de ello es que se necesite recoger más datos. Esta actividad podría aparecer al final de la actividad de recogida de datos, sin embargo, se ha optado por colocarla en este nivel e incluir más tareas de análisis de datos que tienen que esperar a que se recopilen todos los datos.

Para concretar los filtros a emplear en una calibración basada en juicios de expertos el desarrollador puede emplear o inspirarse en la familia de criterios C.it y C.ex descritos en la sección 6.3.1, y para una calibración 3PL-TRI los descritos en la sección 7.3.1.

9.1.3.2. Estimación de los rasgos

Los estadísticos a aplicar a la muestra depurada, así como el orden de aplicación de los mismos durante la **estimación de los parámetros**, dependen de la calibración específica a realizar y difieren, según la experiencia, de la CE a la CT. En ambos casos primeramente se concretan los procedimientos de estimación de los parámetros y seguidamente se obtienen los valores de los parámetros para cada uno de los ítems (Figura 61).

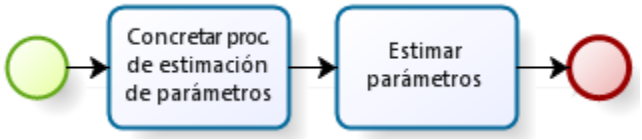


Figura 61.- Detalle del proceso de negocio "Obtener parámetros"

En el caso de la calibración con expertos, y con respecto a los estadísticos a emplear, estos dependerán directamente del tipo de los valores de la variable medida (o del parámetro a estimar). Así, por ejemplo, cuando la variable es escalar (como la dificultad de los ítems), puede emplearse un estadístico similar a M.dif, y cuando la variable es categórica (como la destreza lingüística trabajada por el ítem), puede emplearse un estadístico similar a M.est. La descripción específica de ambos estadísticos se halla en las secciones 6.3.2 y 6.3.3, respectivamente, así como el razonamiento seguido hasta su consecución.

En el caso de la estadística, establecido el modelo de la TRI a utilizar, existen diversos procedimientos alternativos para determinar los parámetros del modelo como son, por ejemplo, las técnicas de máxima verosimilitud (condicionada, conjunta o marginal) o bien las estimaciones bayesianas (marginal, modal o conjunta). En la actualidad existe software específico que computan algunos de los procedimientos mencionados. Por ejemplo, el método de estimación bayesiana MAP aplicado en CT ha sido calculado por XCALIBRE 1.10 para Windows (sección 7.3.2).

9.1.3.3. Análisis posteriores

Una vez establecidos por cada ítem los valores que lo caracterizan, es necesario realizar algunos análisis que avalen los procedimientos de medición empleados y los resultados obtenidos.

En el caso de la calibración basada en expertos, tradicionalmente se ha reconocido una fuente importante de error en la medición de la variabilidad entre observadores (Fleiss, 1986; Landis y Koch, 1977). Consecuentemente, uno de los objetivos de los estudios consiste en estimar el grado de dicha variabilidad. En este sentido, hay dos aspectos distintos que forman parte típicamente del estudio de fiabilidad: por una parte, el *sesgo entre observadores* – o bien dicho con menos rigor, la tendencia de un observador a dar consistentemente valores mayores que otro – y de otra parte, la *concordancia entre observadores (confiabilidad)* – es decir, hasta qué punto los observadores coinciden en su medición. El análisis de fiabilidad de la calibración basada en expertos incluirá el estudio de ambos aspectos (actividad **Analizar fiabilidad del proceso** desarrollado en la Figura 62).

En el caso de la calibración TRI, se realizarán el resto de comprobaciones que quedan pendientes para verificar la adecuación del modelo estadístico empleado a los datos de la muestra (actividad **Comprobar ajuste de datos al modelo** en la Figura 62). La comprobación del ajuste es necesaria para dotar de valor psicométrico a los ítems (López Pina, 1995). En concreto, las comprobaciones que quedan pendientes de realizar en este punto son *la invarianza de los parámetros* y la *independencia local de los ítems*. Obsérvese que en la mayoría de los casos el software empleado para obtener los valores de los parámetros de los ítems genera simultáneamente valores de índices de confiabilidad que pueden usarse para determinar el ajuste de los datos al modelo empleado.

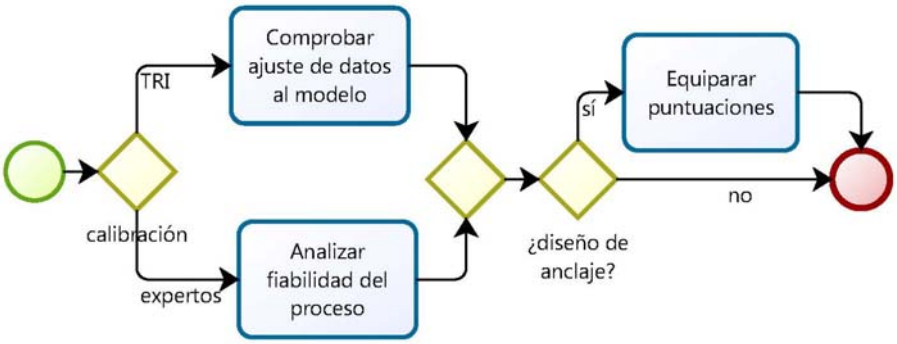


Figura 62.- Detalle del proceso de negocio "Hacer análisis posteriores"

Finalmente, y si ha habido diseño de anclaje, se procederá a unificar las puntuaciones de los distintos subtests establecidas independientemente, para poder compararlas en una escala de

habilidad con origen y unidad comunes. Este proceso que se realizará en la actividad identificada por **Equiparar puntuaciones** de la Figura 62.

En el caso de la calibración CE, para verificar la validez de los resultados, por ejemplo, se emplearon índices Kappa, intervalos de confianza y porcentuales entre los valores de las estimaciones generadas entre los expertos de los grupos PE1 y PE2 consultados. Para la equiparación de puntuaciones en la CT se empleó el método de transformación lineal media-sigma, si bien hay otras alternativas como las propuestas en la sección 5.4.2.2.

PARTE 4
Conclusiones y líneas abiertas

*La **Parte 4** concluye la exposición de la memoria. En ella se exponen las conclusiones alcanzadas mediante el trabajo desarrollado, se indican las aportaciones para la comunidad científica y se comentan posibles líneas de desarrollo a seguir. Así mismo, se resumen las publicaciones que se han generado hasta el momento como consecuencia del trabajo realizado.*

Capítulo 10
Conclusiones y líneas futuras

La línea de mejora de Hezinet, que integra TAIs dentro del sistema para valorar el nivel de conocimientos de un alumno nuevo, se va a convertir en uno de los ejes principales de la continuidad de su éxito. Para empezar, la utilización del sistema a través de Internet permite, además, realizar las pruebas de nivel incluso a distancia. Así mismo, los alumnos tienen una actitud positiva hacia los test informatizados, sean adaptativos o no. Por lo demás, la propia informatización permite que las condiciones de realización de la prueba sean homogéneas para todos los alumnos. También ofrece un rápido procesamiento de las respuestas que se proporcionan, lo que finalmente redunda en un menor tiempo de aplicación, la posibilidad de obtener las puntuaciones al momento y recibir *feedback* inmediato durante la revisión de las respuestas dadas (Olea y Hontangas, 1999)

Los profesores se benefician de que se registra más información sobre las condiciones de realización de la prueba (como, por ejemplo, el tiempo invertido en obtener cada una de las respuestas). Del mismo modo, la automatización de la corrección reduce los errores de corrección de los tests. Si además la prueba es adaptativa, como se eligen los ítems más adecuados según las respuestas aportadas por el alumno, se obtiene más precisión (i.e. menos error) en la medida de la habilidad que con los tests convencionales. Los TAIs también reducen el tiempo necesario para obtener una valoración, ya que consiguen niveles similares de precisión que los tests convencionales con un número menor de ítems, lo que implica

que se pueden atender más demandas de alumnos mismo tiempo. Finalmente, no se hace necesaria la presencia de una persona para obtener las valoraciones de las pruebas, de manera que el tiempo de atención puede ser más amplio (hasta podría llegar incluso a 24x7). Y como en los TAIs los alumnos no tienen por qué responder a un grupo de ítems prefijado de antemano, se mejora la seguridad de la prueba, ya que el conocimiento de la existencia de un ítem no significa necesariamente que el alumno tenga que responder al mismo (Olea y Ponsoda, 2003).

Para los creadores de material, la posibilidad de realizar los test automatizados permite establecer controles para mejorar la seguridad de la prueba (por ejemplo, conocer el grado de uso de cada uno de los ítems). Además, el hecho ya comentado de que no se utilicen los mismos ítems para evaluar el nivel de cada uno de los alumnos, aunque no significa que se resuelva el problema de la copia o transmisión de los ítems (suele ocurrir que un escaso porcentaje de ítems se aplican a muchos alumnos), el problema es menor que en los tests convencionales de lápiz y papel o en los tests informatizados fijos.

Como se comenta en la introducción, esta tesis tiene tres objetivos que se han cumplido. Primeramente, se ha determinado empíricamente que existen dos métodos posibles para la calibración de ítems que generan estimaciones equiparables y se ha propuesto un proceso de negocio para realizarlo, tanto en el caso de seguir la TRI como si se quieren utilizar juicios de expertos para decidir los niveles de dificultad.

Además, y basándose en el desarrollo empírico, se ha formalizado un proceso de negocio para cada uno de los procesos de calibración. Actualmente se está trabajando en la construcción de un sistema de ayuda a la toma de decisiones durante la calibración basado en la instanciación del proceso de negocio de calibración en cualquiera de las dos versiones desarrolladas. Además se han establecido una serie de indicadores para poder comparar el coste de recursos que supone cada uno de ellos.

En las siguientes secciones se exponen las conclusiones del trabajo realizado, las aportaciones que ha generado éste a la comunidad

científica, las líneas futuras de trabajo identificadas y las publicaciones realizadas.

10.1. Conclusiones

La conclusión fundamental de este trabajo es que ***una calibración de ítems siguiendo la TRI es equiparable a una calibración con expertos*** cuando se consideran los juicios de 7 o más expertos. Para demostrarlo se ha calibrado un banco de 252 ítems que estaban siendo utilizados para determinar el nivel de conocimiento de euskara. Se han creado estadísticos y filtros propios para realizar la comparación, tal y como se comenta en detalle en el Capítulo 8.

Se ha creado **el primer banco de ítems de euskara calibrado utilizando la Teoría de Respuesta al Ítem** para el aprendizaje del idioma. Esta calibración ha dado origen también al primer Test Adaptativo Informatizado de euskara utilizado para la clasificación de estudiantes nuevos. Este trabajo forma parte de la tesis del Dr. Lopez Cuadrado (López-Cuadrado, 2008) y se documenta en el Capítulo 7.

Así mismo, el banco de ítems también es el primero que se ha **calibrado basándose en juicios de múltiples expertos.** En esta ocasión, se han utilizado estadísticos para la medición de los rasgos concretos y se ha formalizado un procedimiento concreto para su realización. Este trabajo será la base de la tesis de la profesora Dña. Ana Jesús Armendariz y se comenta en el Capítulo 6 de esta memoria.

Como resultado de ambas calibraciones, se observa que **el banco de ítems de partida, está sesgado y no cubre todos los niveles**. Se nota una desviación hacia los niveles de dificultad bajos, observándose escasez de ítems que cubran los niveles altos. Este problema, que se debe resolver añadiendo nuevos ítems, no ha resultado problemático hasta el momento porque la mayoría de los alumnos que usaban el sistema correspondía a los primeros niveles. Si no se corrigiera, el efecto del TAI sería contraproducente, ya que los ítems que se podrían administrar al alumno no serían válidos para aportar información sobre el nivel del mismo, llevando al

sistema a no poder certificar su nivel con un nivel de confianza suficiente al alcanzarse el número máximo de ítems por test.

Atendiendo a los costes que suponen las calibraciones, se establecen como indicadores del consumo de recursos de cada uno de los procesos el *tiempo invertido,* el *número de valoraciones a recoger* y *el número de sujetos activos y pasivos que hay que involucrar*. Hay un segundo nivel hay más indicadores relacionados con los anteriores (por ejemplo, la *tasa de abandono de participantes* influye en el número de sujetos pasivos que hay que involucrar). Todos estos aspectos se comentan en el Capítulo 8.

A partir de los indicadores definidos se concluye que **cuanto mayor sea el banco de ítems, mayor será el coste de calibración.** También que **el uso de métodos síncronos alarga el tiempo de realización de los procesos de calibración** ya que se dan tiempos de espera para conseguir la sincronización de las partes. Finalmente, se deduce que, en cualquier caso, **la calibración mediante juicios de expertos es menos costosa que la de TRI,** ya que el número de sujetos a involucrar es sensiblemente menor y el consumo de recursos asociados, también. Esta comparación se comenta en el Capítulo 8.

De los experimentos realizados se sugiere que, **ante una disposición limitada de recursos se utilice la calibración mediante juicios de expertos,** que puede obtener unos mínimos resultados con menos recursos. Sin embargo, las pruebas estadísticas que ofrece la TRI pueden dar lugar a estimaciones más completas (de más parámetros) que pueden también conseguir un afinamiento más preciso del funcionamiento del sistema, por lo que, se recomienda siempre que se tenga facilidad para disponer de los recursos correspondientes.

10.2. Aportaciones principales

Del trabajo de investigación realizado, se pueden recoger múltiples aportaciones, que se documentan en los siguientes párrafos de este apartado.

En primer lugar, **se ha propuesto un proceso de negocio para desarrollar calibraciones de ítems basadas en juicios de**

expertos. Hasta el momento, existían diferentes alternativas para calibrar los ítems utilizando la TRI, pero no se contemplaba ningún otro método de estimación de las dificultades de los ítems. Se ha comprobado que la propuesta es válida, ya que experimentalmente se ha seguido el proceso de negocio en los experimentos PE1, PE2 y CE, tal y como se comenta en el Capítulo 6 de esta memoria.

La **calibración de los ítems se ha ejecutado aplicando la técnica Plan-Do-Check-Act de calidad**: identificando las tareas a realizar, y una vez ejecutadas, buscando oportunidades de mejora de las mismas para posteriores instanciaciones. Tan solo es necesaria un aumento de la supervisión de las tareas que se realizan, ya que hacen falta datos sobre cómo se ha desarrollado la ejecución de las mismas. En este caso no suponían mayor esfuerzo, ya que se quería obtener información sobre los recursos que suponía cada una de las aproximaciones estudiadas.

Aunque el procedimiento de la calibración según la TRI está documentado en la bibliografía existente, no ocurre lo mismo para la calibración basada en juicios de expertos. El que sigue la TRI tiene varias alternativas posibles que no se han implementado, ya que, como se comenta en el Capítulo 1, no se considera objetivo de esta tesis. El proceso de calibración mediante juicios de expertos es original. El **proceso de negocio se ha formalizado utilizando BPMN.** Es un **proceso de negocio parametrizado.** Los diferentes parámetros *(tamaño del banco de ítems, longitud de los cuestionarios, número de participantes, plazos, tasas previstas de abandono, tasas previstas de descarte)*. La propuesta de proceso se puede encontrar en el Capítulo 9. En el anexo 7 se adjunta una breve descripción de BPMN que puede ayudar a los noveles en el uso de esa notación.

El **modelo estadístico de la calibración con expertos es original e incluye filtros** para validar los datos muestrales **y estimadores** para determinar de forma *off-line* los valores más probables entre los pronósticos más consensuados sobre los rasgos considerados. En concreto, la calibración de los ítems del experimento CE se ha realizado tanto para variables continuas (rasgo *dificultad*) como para variables discretas (rasgo *destreza*).

Utilizando los experimentos que demuestran que el proceso propuesto es viable, se ha hecho una **estimación de recursos**

necesarios para hacer una calibración de ítems. Se ofrecen indicadores que muestran la evolución en los costes de los procesos e, implícitamente, se dan pautas para la inclusión de otros indicadores dentro de los procesos de negocio. Ese estudio de costes se puede encontrar en el Capítulo 8.

Se han sentado las **bases para la creación del sistema experto CALLIE,** que ayudará en las tareas de calibración y de cuyo desarrollo se encargará la doctoranda Ana Jesús Armendariz. La identificación de los procesos de negocio posibilita acometer mejoras futuras y puede servir de ejemplo para otros procesos. Concretamente, y puesto que ambos procesos de negocio propuestos comparten la misma arquitectura principal, sería viable construir una única herramienta software que, dependiendo de las necesidades específicas de los desarrolladores en cada ocasión, asistiera en la construcción de una calibración de ítems eficaz y eficiente.

10.3. Líneas futuras

El trabajo aquí presentado deja abiertas líneas de trabajo que se abordarán en el futuro. En los siguientes párrafos se describen las mismas.

En primer lugar, habría que investigar **cuál es el número mínimo de expertos con el que se consigue que exista correlación entre los dos métodos de calibración comparados**. El trabajo desarrollado en esta tesis nos dice que con siete o más expertos se pueden equiparar las calibraciones de ítems. Se ha tomado el número de otras áreas como la *evaluación de sistemas de* software o la *evaluación de interfaces,* en las que también se recopila información de expertos. Sin embargo, también hemos encontrado algún trabajo de investigación que indica que un número menor de expertos no es suficiente para realizar la calibración (Conejo, Guzmám et al., 2008) y los datos recopilados nos permitirán realizar este tipo de análisis.

Con la muestra de los datos recogida se pueden hacer otro análisis de datos pendientes. Por ejemplo, hay datos, como los perfiles de los expertos, los tiempos utilizados por ellos, etc. que aún no han sido analizados. También hay algunas alternativas que se han

pospuesto a la finalización de la defensa de esta tesis. Por ejemplo, Puesto que con los juicios de expertos solo se estima un rasgo, con los datos de la CT se podría **efectuar** otra calibración alternativa empleando el modelo de Rasch de un parámetro logístico (**1PL-TRI**). La calibración resultante podría **compararse con las** ya **obtenidas** y el intervalo de confianza de la estimación sería mayor al ser necesarias menos administraciones.

Se ha comenzado a trabajar con programas de simulación para **verificar que a los alumnos nuevos se les adjudica el nivel de conocimiento que tienen.** Se trata de garantizar que el sistema TAI cumple con sus objetivos. Sobre todo, es necesario el afinamiento sobre la decisión del estrato concreto de un curso al que tiene que ir un nuevo alumno.

El estudio de costes se completará con **fórmulas parametrizadas para prever el coste** que tendría una calibración específica antes de realizarla los modelos de calibración propuestos. Los experimentos realizados se reenunciarán como **una familia de experimentos controlados** siguiendo el enfoque *Goal/Question/Metric* como en (Wohlin, Runeson et al. 2000; Otero and Dolado 2007; Rolón, García et al. 2007). Esta formulación es de interés para investigadores del área de la Ingeniería del Software empírica cuyo trabajo está focalizado en la medición y evaluación de productos software en lugar de la producción directa de software.

Como se ha indicado previamente, el sistema CALLIE implementará los procesos de negocio que se describen en esta memoria. Concretamente, se incluirán dentro de un **sistema experto para la calibración de ítems** que guiará al usuario de manera sencilla y sin exigirle conocimientos específicos en psicometría ni estadística, para efectuar las calibraciones de bancos de ítems que hemos visto en esta memoria (López-Cuadrado, Armendariz et al., 2008). La herramienta será un módulo autónomo e integrable en Hezinet.

Relacionado con la construcción de una única herramienta que integre ambas calibraciones, sería interesante añadir a ambos procesos de negocio una última actividad de **conversión de escala de dificultad** de los ítems calibrados. Con esta nueva opción y con independencia del proceso de calibración escogido, la dificultad de

los ítems quedaría estimada siempre en **dos escalas alternativas**: la estadística con rango (-∞, +∞) y 0 como nivel medio de dificultad y la de los expertos, que por ejemplo, para Hezinet específicamente es [1, 12]. Esta opción permitiría incorporar nuevos bancos calibrados estadísticamente a bancos existentes y calibrados mediante expertos, de manera que algoritmo seleccionador de ítems del sistema de conocimiento focalizaría su atención en el valor convertido en lugar del estimado. Y también a la inversa, añadir bancos cuyos ítems originariamente se han estimado mediante valoraciones a sistemas que emplean ya ítems con estimaciones estadísticas (López-Cuadrado, Armendariz et al., 2009).

La administración de tests a distintos colectivos ha hecho aflorar la necesidad de realizar informes a partir de los datos de las distintas administraciones. Así, de la realización de la PT2, se han tenido que construir distintos **informes con los resultados**. Estos informes, podrían estar automatizados en su mayor parte, ofreciendo resultados estadísticos de los resultados para poder ser aprovechados por las instituciones que han colaborado activamente en la administración de tests. En la actualidad se está trabajando para que el sistema de administración electrónica de subtests, realice estos informes de manera automática y reduzca el trabajo de análisis por parte de los profesionales encargados de hacer las pruebas.

10.4. Publicaciones generadas

En esta sección se muestran las distintas publicaciones donde se recogen las aportaciones principales de esta tesis. En total se ha generado 1 capítulo de libro (en inglés), 9 artículos en congresos internacionales, 4 más en congresos nacionales y 10 informes técnicos. Éstas son:

Capítulo de libro

- López-Cuadrado, J., A. J. Armendariz, T. A. Pérez, R. Arruabarrena y J. A. Vadillo (2009). Chapter "Computerized adaptive testing, the item bank calibration, and a tool for easing the process." International Technology, Education and Development Conference: 1-22. ISBN 978-953-7619-40-4.

En esta publicación se identifican y se caracterizan los módulos que componen el sistema CALLIE para la ayuda en la toma de decisiones en procesos de calibración de ítems. Además, el sistema ayuda en la gestión de las muestras de los sujetos administrados, en aspectos psicométricos y en los cálculos estadísticos a realizar

Publicaciones en congresos internacionales

- López-Cuadrado, J., A. J. Armendariz, T. A. Pérez y R. Arruabarrena (2008). Helping Tools For Item Bank Calibration And Development Of Computrized Adaptive Tests. En L. G. Chova, D. M. Belenguer y I. C. Torres (eds.) Proc. of International Technology, Education and Development Conference (INTED'08), Valencia (España), International Association of Technology, Education And Development (IATED): 1-9.

  Esta publicación presenta un prototipo denominado CALLIE para la ayuda en la toma de decisiones durante los procesos de calibraciones de ítems. El prototipo está basado en los modelos de negocios para calibraciones de expertos y 3PL-TRI propuestos en esta memoria de tesis.

- Arruabarrena, R. y J. López-Cuadrado (2006). Issues to be taken into account when calibrating items. En A. Méndez-Vilas, A. Solano, J. M. González y J. A. M. González (eds.) Proc. of Current Developments in Technology-Assisted Education, Sevilla (España), Formatex Research Center-Badajoz, 2: 906-910.

  En ella se analizan y comparan los costes de realizar las calibraciones CE y la CT.

- Arruabarrena, R., S. Sanz-Santamaría y T. A. Pérez (2005). Quality Techniques Help in LMS' Improvement. En A. Méndez-Vilas, B. González-Pereira, J. M. González y J. A. M. González (eds.) Proc. of Recent Research Developments in Learning Technologies, Cáceres (Spain), Recent Research Developments in Learning Technologies, Formatex Reseach Center, **1**: 199-203.

  En esta publicación se describe cómo ha beneficiado el uso de la técnica PDCA a la hora de identificar y mejorar procesos de

captación de información, haciendo posible la ejecución posterior de réplicas mejoradas.

- Arruabarrena, R., J. A. Vadillo y J. Gutiérrez (2003). Are Experts Difficulty Guessing And Statistical Results Comparable? En A. Méndez-Vilas, J. A. M. González y J. M. González (eds.) Proc. of Advances in technology-based education: towards a knowledge-based society, Badajoz (España), Sociedad de la Información, Junta de Extremadura (CECT), **1**: 542-545.

  En ella se presentó el plan a seguir para determinar si las dos líneas de calibración de ítems más comunes, la basada en aportaciones de expertos y la determinada por algún método estadístico, son o no equiparables.

- Arruabarrena, R., T. A. Pérez, J. Gutiérrez, J. López-Cuadrado y J. A. Vadillo (2002). On Evaluating Adaptive Systems for Education. En P. D. Bra, P. Brusilovsky y R. Conejo (eds.) Proc. of AH2002, 2nd. International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Málaga, Lecture Notes in Computer Science (2347), Springer-Verlag,: 363-367.

  En ella se presenta una recopilación de técnicas de evaluación de sistemas educativos. Incluye un plan de evaluación para mejorar Hezinet que detalla los tipos y técnicas de evaluación a emplear.

- López-Cuadrado, J., T. A. Pérez, J. A. Vadillo y R. Arruabarrena (2002). Integrating Adaptive Testing in an Educational System. En E. Kähkönen y E. Sutinen (eds.) Proc. of Educational Technology in Cultural Context: ETCC2002. First International Conference on, Joensuu, Finland, (Joensuun Yliopisto, Intenational Proceeding Series), University of Joensuu: 133-139.

  En ella se aborda la conveniencia de que los tests, que realizan los alumnos para medir la evolución de su aprendizaje en sistemas educativos informatizados, sean adaptativos. Ello implica que, para que el compilador de tests genere un test ad-hoc para un alumno específico, los ítems del sistema deben estar calibrados.

- Villamañe, M., J. Gutiérrez, R. Arruabarrena, T. A. Pérez, S. Sanz-Lumbier, S. Sanz-Santamaría y J. López-Cuadrado (2001). Use and Evaluation of HEZINET: a system for Basque language

learning. En (eds.) Proc. of ICCE, Seoul, South Korea, AACE: 93-101.

En esta publicación se presentan los resultados de una evaluación específica realizada a Hezinet. La evaluación de la aplicación se centra en la eficacia instruccional del sistema y su adaptación automática con respecto a los usuarios.

- Arruabarrena, R., T. A. Pérez, J. Gutiérrez, J. López-Cuadrado y J. A. Vadillo (2001). Compendio de técnicas para evaluación de sistemas hipermedia adaptativos. En (eds.) Proc. of Simposium Internacional de Informática Educativa, SIIE, Instituto Superior Politécnico de Viseu, Viseu, (Portugal), Escola Superior de Educaciçao: 10.

  En esta publicación se ha presenta un compendio de técnicas de evaluación de sistemas educativos. Estas técnicas han sido extraídas principalmente de sistemas hipermedias adaptativos.

- López-Cuadrado, J., T. A. Pérez, R. Arruabarrena, J. A. Vadillo y J. Gutiérrez (2002). Generation of Computerized Adaptive Tests in an Adaptive Hypermedia System. En A. Méndez Vilas, J. A. Mesa González y I. Solo de Zaldívar Maldonado (eds.) Proc. of Educational technology - Information society and education: monitoring a revolution, Badajoz (Spain), Sociedad de la Información, Junta de Extremadura (CECT), **2**: 674-678.

  En esta publicación se aborda cómo la TRI puede ser empleada para añadir nuevas funcionalidades al sistema Hezinet. Entre las mejoras identificadas están la incorporación de tres módulos: uno para compilar tests adaptativos informatizados y los otros dos para generar calibraciones 3PL-TRI de ítems existentes de forma off-line y de forma on-line.

Publicaciones en congresos nacionales

- Arruabarrena, R., S. Sanz-Santamaría y J. Gutiérrez (2007b). Desarrollo eficiente de pruebas de campo. En I. F. d. Castro (eds.) Proc. of Simposio Nacional de Tecnologías de la Información y las Comunicaciones en la Educación: Sintice-2007, incluido en el II Congreso Español De Informática: CEDI'07 (SINTICE-CEDI'07), Zaragoza (España), Nuevos retos

científicos y tecnológicos en Ingeniería Informática, Thomsom, **1**: 309-312.

En esta publicación se describe la aplicación de técnicas de calidad para reducir costes en las pruebas PE1 y PE2.

- Arruabarrena, R. y T. A. Pérez (2005c). Una experiencia arbitrando incidencias producidas en pruebas de campo. En M. O. Cantero (eds.) Proc. of VI congreso nacional de Informática Educativa. I Simposio Nacional de Tecnologías de la Información y las Comunicaciones en la Educación: Sintice-2005 (SINTICE-CEDI'05), Granada, Thomson Paraninfo (Spain): 161-166.

  En ella se enumeran los problemas afrontados a la hora de desarrollar las pruebas de campo con expertos así como las lecciones aprendidas.

- López-Cuadrado, J., M. Villamañe, J. Gutiérrez, T. A. Pérez, R. Arruabarrena y J. A. Vadillo (2001). CiberBiblio: una biblioteca virtual multimedia. En P. d. l. F. R. y. A. P. Alarcón (eds.) Proc. of II Jornadas Españolas de Bibliotecas Digitales: JIBIDI'01, Almagro, Ciudad Real (España): 12.

  En esta publicación se presenta una biblioteca digital de recursos multimedia interactivos que incluye elementos dedicados al fomento del euskera, entre ellos, la plataforma Hezinet para el aprendizaje del euskera y el servidor web de revistas y periódicos digitales *Kiosk@*.

- Gutiérrez, J., T. A. Pérez, R. Arruabarrena y J. López-Cuadrado (2001). Evaluación en Sistemas Hipermedia Adaptativos. En P. M. Hontangas (eds.) Proc. of Congreso de Metodología de las Ciencias Sociales y de la Salud, Madrid, Simposium sobre tests adaptativos de enseñanza y diagnóstico informatizados, Asociación Española de Metodología y Ciencias del Comportamiento, AEMCCO:

  En esta publicación se presentaron por primera vez ideas acerca de la incorporación a Hezinet de las técnicas de adaptación utilizando la TRI.

Informes técnicos

Relacionados con ambas calibraciones:

- Arruabarrena, R., J. López-Cuadrado y A. J. Armendariz (2007a). Consideraciones para el cómputo de costes de calibraciones de bancos de ítems. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 06-2007): 37.
- Arruabarrena, R. (2005). Filtrado de un banco de ítems. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 02-2005): 60.

Relacionados con la calibración con expertos:

- Arruabarrena, R. y T. A. Pérez (2010). Calibración de ítems con expertos: procesos BPM, ejecución, análisis y mejora. Una investigación empírica. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 08-2010): 53.
- Arruabarrena, R. y A. J. Armendariz (2008). Estimación de los parámetros de los ítems de un sistema de e-learning vía expertos. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 02-2008): 244.
- Arruabarrena, R. y T. A. Pérez (2005a). Arbitraje de las incidencias producidas en pruebas de campo para calibrar un banco de ítems con expertos. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 09-2005): 9.
- Arruabarrena, R. y T. A. Pérez (2005b). Pruebas de campo para calibrar un banco de ítems vía expertos. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 08-2005): 96.

Relacionados con la calibración estadística:

- Arruabarrena, R., López-Cuadrado, J. y A. J. Armendariz (2010). Calibración 3PL-TRI de ítems: procesos BPM, ejecución, análisis y mejora. Una investigación empírica. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 09-2010): 44.
- López-Cuadrado, J. and R. Arruabarrena. Diseño de anclaje de un banco de ítems. San Sebastián, University of the Basque Country. Technical report. UPV/EHU/LSI/TR 12-2005. p. 109.
- López-Cuadrado, J., R. Arruabarrena y A. J. Armendariz (2005a). Aita Larramendi Andoaingo Ikastolako ikasleen euskara frogen emaitzen txostena. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 24-2005): 22.

- López-Cuadrado, J., R. Arruabarrena y A. J. Armendariz (2005b). La Salle Andoaingo Ikastolako ikasleen euskara frogen emaitzen txostena. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 25-2005): 40.

PARTE 5
Anexos y bibliografía

El anexo ***1 Banco de ítems administrados*** *recoge el enunciado y las opciones de respuestas de los ítems administrados en las pruebas de campo conducidas.*

El anexo ***2 Cuestionario de la CE*** *muestra íntegramente uno de los cuestionarios administrados a los expertos, específicamente, el cuestionario número 8 de la PE1.*

El anexo ***3 Resultados de la calibración CE*** *recoge las estimaciones de dificultad y destreza estimadas a partir de las valoraciones de los expertos. También muestra las frecuencias específicas de éstas por ítem una vez depura la muestra, y los valores de los análisis de fiabilidad realizados para salvaguardar la validez de los resultados.*

El anexo ***4 Cuestionario electrónico de la CT*** *presenta las pantallas más significativas de los cuestionarios electrónicos, así como los diversos correos electrónicos remitidos durante el desarrollo de las pruebas no supervisadas, las PT1.*

El anexo ***5 Resultados de la calibración 3PL-TRI*** *muestra el resultado final de la calibración del banco de ítems según el modelo 3PL de la TRI. Para los ítems que prevalecen en el banco se muestran los valores estimados una vez equiparadas las puntuaciones, y para el resto el motivo por el cual han sido descartados.*

El anexo ***6 Valores de contraste: CE vs CT*** *recoge los resultados de los contrastes de la prueba de Wilcoxon y del T-test empleados en la evaluación multicriterio. Contiene también los valores muestrales (los estimados) y los normalizados, a los cuales se han aplicado estas pruebas. Además, se incluyen las extrapolaciones de los costes temporales y económicos de la calibración CE (CT, respectivamente) para tamaños de bancos de ítems alternativos empleando solo aportaciones recabadas a través de pruebas PE2 (PT2, respectivamente).*

El anexo ***7 Póster BPMN 1.1*** *incluye como su nombre bien indica un póster que presenta de forma sintetizada los elementos disponibles en la notación estándar BPM 1.1. En concreto, BPMN define la notación y la semántica de un Diagrama de Procesos de Negocio. Incluye varios diagramas ilustrativos.*

La sección dedicada a las ***Referencias bibliográficas*** *recoge la bibliografía utilizada y referida, para que las personas interesadas puedan profundizar en los aspectos tratados.*

ANEXO 1 Banco de ítems administrados

La siguiente tabla recoge los 252 ítems del banco que se han administrado tanto a expertos como a sujetos anónimos, con vistas a construir dos calibraciones independientes del mismo banco de ítems, una basada en los juicios de los expertos y otra estadística a partir de la muestra de sujetos anónimos.

El banco original fue proporcionado por los productores de la fundación Aurten Bai/Zornotzako Barnetegia. Tras un filtrado inicial por parte de expertos en la materia para, principalmente, actualizar los ítems a las últimas normas de lingüísticas del euskera y uniformizar estilos, el banco actualizado resultante es el recogido en la siguiente tabla.

Su estructura la definen un identificador para referenciar cada uno de los ítems (columna "*ID*"), el enunciado del ítem (columna "Enunciado"), y las cuatro opciones de respuesta (columnas "*Opción1*", "*Opción2*", "*Opción3*" *y* "*Opción4*"*)*.

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
1	Zuek zer zarete?	Gu ikasleak gara.	Zuek ikasleak zarete.	Gu ikasleak dira.	Haiek ikasleak dira.
2	Norena da liburua?	Liburua urdina da.	Liburua etxean da.	Nirea da liburua.	Nire liburua da.
3	Non dago Urgull mendia?	Urgull mendia han dago.	Urgull mendia oso zikin dago.	Urgull mendia erreta dago.	Urgull mendia polita dago.
4	Gu orain Donostian gaude.	Zuek orain zer zabiltzate?	Nor dago orain Donostian?	Zuek Donostian bizi al zarete?	Zuek orain non zaudete?
5	Urgull mendia handia da.	Urgull mendia zein da?	Urgull mendia nolakoa da?	Urgull mendia oso altua da?	Urgull mendia nola txikia da?
6	Nongoak zarete zuek?	Bilbotarrak gara gu.	Donostiakoak gaude gu.	Euskalerrian gara gu.	Oso ederrak gara gu.
7	Zuek Donostian bizi al zarete?	Ez, gu ez bizi gara Donostian.	Donostiakoak gaude gu.	Ez, gu ez gara Donostian bizi.	Ez, ez bizi gu gara Donostian.
8	Nork ditu nire liburuak?	Ni ditut zure liburuak.	Haiek dute zure liburuak.	Haiek dituzte zure liburuak.	Haiek daukazu zure liburuak.
9	Bizkaiko ..... goaz.	mendiara	mendietara	menditara	mendia
10	..... habietan usoak daude.	Zuhaitz horietarako	Zuhaitz horietako	Zuhaitz hauetarako	Horietako zuhaitz
11	Noren euritakoa da hori?	Euritako hori nirea da.	Euritako nirea da hori.	Hori nire euritakoa da.	Hori euritakoa da nirea.
12	Nire lagunak goizean Bilbon egon dira.	Zer egin dute zure lagunek goizean Bilbon?	Non egon dira zure lagunak goizean?	Nortzuk egon dira goizean Bilbon?	Noiz egon dira zure lagunak Bilbon?
13	Nora doaz txirrindulariak?	Zure etxetara.	Edozein lekura.	Lasterketa horretara.	Arratsaldeko lehiaketatara.
14	Museoa ikusgarria da. Antzokia ikusgarria da.	Museoa ikusgarria da, antzokia ere.	Museoa ikusgarria da baita antzokia ere bai.	Museoa ikusgarria da, antzokia ere bai.	Museoa ikusgarria da, antzokia ere da.
15	Non utzi duzu liburua?	Ohe gainean.	Mahaiaren artean.	Mahaian azpian.	Hor, Ixabel ondoan.
16	Gazta horiek baserri horietan egin dituzte.	Gaztak baserrikoak dira.	Gaztak baserrikoa dira.	Gaztak baserrietakoak dira.	Gaztak baserrietako dira.
17	..... ikusi	Ikasten	Ikastea	Ikasteko	Ikastera

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	ditut.				
18	Nola idazten da hizkiz '17:30'?	Bostak eta erdiak dira.	Bostak eta erdi dira.	Bost eta erdiak dira.	Bost eta erdi dira.
19	Aktoreek ez ..... saririk irabazi.	dituzte	dute	ditu	du
20	..... atzamarra harrapatu du.	Atea ixteko	Atea ixtera	Atea ixten	Atea ixtean
21	Telebista ikusten ari al zara?	Bai, telebista ikusten naiz.	Bai, telebista naiz ikusten ari.	Bai, telebista ikusten ari naiz.	Bai, ikusten naiz telebista ari.
22	Irratia entzuten ari al zarete?	Ez, ez gara irratia entzuten ari.	Ez, ez ari gara irratia entzuten.	Ez, ez gara irratia ari entzuten.	Ez gara irratia ahal entzuten ari.
23	Nora joango zarete bihar?	Bihar Lezora joango naiz.	Bihar Elizondora joaten gara.	Bihar Elizondora goaz.	Bihar Baionara joango gara.
24	Egunero idazten dute orri bat.	Zer egiten dute egunero?	Nola idazten dute orri bat?	Nondik idazten duzue orri bat?	Noiz idazten dute orri bat?
25	Lehen egunkaria erosi dut.	Nork erosi du egunkaria?	Noiz erosi duzu egunkaria?	Lehen zer erosi duzu?	Egunkaria erosi al duzu?
26	Zein da zuzena?	Ez dute ezer esan.	Ez dute ezer ez esan.	Ez dute zerbait esan.	Ezer esan ez dute.
27	Alkandora ..... 2.000 pezeta balio du.	hau	honek	hauek	hauk
28	Gero filme bat ikustera joango naiz. Oraintxe	joaten dut.	joaten ari naiz.	joaten naiz.	noa.
29	Norentzat erosi duzu oparia? ..... erosi dut.	Umerentzat	Umentzat	Umearentzat	Umeentzako
30	Dendan ..... ikusiko dut.	sartzean	sartuan	sartzeko	sartzenean
31	Zerezkoak dira euskaltegiko ateak?	Egurrezkoak dira.	Egurrez dira.	Egurrezko dira.	Egurrezko eginda daude.
32	Tolosara	joan aurretik	joan lehen	aurrejoan	aurrean joan

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	..... aldatu genuen gurpila.		baino		
33	Zure lehengusuek ez ..... zure izena?	dakizu	dakite	dakizue	daki
34	..... lan egiten du.	Astoek bezala	Astoak bezalakoa	Bezala astoek	Astoa bezala
35	..... hurbildu da txakurra.	Artzain honetara	Artzain hangana	Artzain honengana	Artzain honi
36	Mendira ..... ekarri mesedez eguzki lorea.	noiz joaten zaren,	joaten noizean,	joaten zarenean,	joaten zaren noizean,
37	"El programa sobre música" gaur da.	Programa gainean musika	Musikari buruzko programa	Programa buruz musika	Programa musikari buruz
38	..... noa, hau da amaiera.	Zerurantz	Zeruan	Zeruaz	Zerutik
39	Zigarro bat nahi al duzu?	Ez, ez nahi dut zigarrorik.	Ez, ez zigarrorik nahi dut.	Ez, ez zigarrorik dut nahi.	Ez, ez dut zigarrorik nahi.
40	Noiztik ez duzu ikusi?	Joan den astetatik.	Abenduaren bostetatik.	Astelehenetik.	Goizeko hamarretik.
41	Ahaztu zait Amerikako osabari	idazteko.	idaztera.	idaztea.	idatziaz.
42	Aurten hori nahi du eta datorren urtean gehiago	nahiko izango du.	nahi izango du.	nahi izango da.	nahi du.
43	Zein ..... ahaztu zaio koadernoa?	ikaslei	ikasleari	ikasleeri	ikasleri
44	"Antzarak ferratzera joan zaitez" esan dio. Zer esan dio?	Antzarak ferratzera joan dela.	Antzarak ferratzera joateko.	Antzarak ferratzera joan zaitezen.	Antzarak ferratzera joan zaitezela.
45	Liburua ikusi dugu. Liburua puskatuta dago.	Liburu ikusi dugun puskatuta dago.	Liburua ikusi dugula puskatuta dago.	Ikusi dugun liburua puskatuta dago.	Liburua ikusi duguna puskatuta dago.
46	Aukeratu	Ez dago inor	Ez da inor ez	Ez dago inor.	Ez dagoen inor.

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	zuzena.	egon.	egoten.		
47	..... da trena hartzeko.	Beranduena	Beranduegi	Berandu gehiegi	Berandu gehiago
48	Matusalen zaharra da. Hori oso oso zaharra da.	Matusalen baino zaharregia da.	Matusalen baino gehiago zaharra da.	Matusalen baino zaharragoa da.	Da gehiago zaharra Matusalen baino.
49	Zuk gu ez ..... ikusten betaurreko horiekin, ez da harritzekoa!	gaituzu	zaitugu	zaituztegu	gaituzue
50	..... edan dut.	Xanpaineko kopa bat	Xanpain kopa bat	Kopa bat xanpain	Bat kopa xanpain
51	Nire prakak ..... dira.	zureak bezalakoak	zureak bezala	zure bezain	zure bezalakoak
52	Anderrek diru gutxi dauka. Ainhoak diru asko dauka.	Anderrek Ainhoak baino diru gehiago dauka.	Anderrek Ainhoak baino diru gutxiago dauka.	Ainhoa Ander baino diruagoa da.	Ainhoak Anderrek baino diru gutxiago dauka.
53	Bai Jeronima bai, maite ..... baina nire bihotza beste batentzat da.	dizut,	zaitut,	zaituztet,	dute,
54	Hori hitzaldirik ..... izan zen.	interesgarria	interesgarriagoa	interesgarriena	interesgarriegia
55	..... tapoia kendu behar zaio.	Ikustea	Ikusteko	Ikusten	Ikustera
56	Sentitzen dut, baina ..... zuekin gero.	ezingo dut joan	ezin izango naiz joan	ezin naiz joan	ezin izan naiz joan
57	Ez kezkatu! Normalean zure lagunak ordu bietarako	heltzen ari dira.	heldu dira.	heltzen dira.	helduko dira.
58	Konzentra zaitez! ....., ez duzu proba	Baina	Bestela	Ordea	Aldiz

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	gaindituko.				
59	Zuek nirekin ....., ez dizuet kontzerturako sarrerarik emango.	ez bazatoz	etorriko ez bazarete	ez bazara etorriko	ez bazatozte
60	Atzerritik datozenak ..... dira.	atzerrikoiak	atzerritakoak	atzerritarrak	atzerrizaleak
61	Jaietan eta, animatzen egoten dena ..... da.	animazalea	animalea	animagaitza	animatzailea
62	Saltzailearen antonimoa ..... da.	eroslea	erosterraza	erostzailea	erostezina
63	Zein dago ondo?	Zapatagilea.	Bronkatzailea.	Jagonlea.	Idazgilea.
64	Ez dakit ..... film horretaz.	gogoratuko naizen	gogoratuko banaiz	gogoratuko naizela	gogoratuko banaizenik
65	Mutilak hori esan du. Mutila ekarri dugu.	Hori esan duen mutilak ekarri dugu.	Hori esan dugun mutilak ekarri du.	Hori esan duen mutila ekarri dugu.	Hori esan dugun mutila ekarri dugu.
66	Ez dut ..... horrelakorik ikusi.	inon ez	non edo non	nonbait	inon
67	..... ez dut itxaropenik.	Politikoetan	Politikoan	Politikoengan	Politikoengana
68	..... jasotako laguntza oso mesedegarria izan da guretzat.	Zuregandik	Zutik	Zuengan	Zurengandik
69	Hori ez diot "a cualquiera" esango.	edozeri	edonori	norbaiti	inori
70	Zorroan ez ezazu eraman	hainbeste dirua.	hain diru.	bezain bat diru.	hainbeste diru.
71	Ez da ..... guda atomikoa izatea.	hain zaila	beste zaila	bezain zaila	zaila bezain

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
72	Zuk niri eskaini ..... liburuak eta zure lehengusuak aurpegi txarra jarri zuen.	zenizkidan	zenizkidaten	nizkizun	zenizkioten
73	Gaixorik egon ginenean, zuk gozokiak ekarri	zenizkigun.	genizkizun.	zenizkiguzun.	dizkiguzun.
74	Nirekin maiteminduta zeundela jakin nuen, egunero begiratzen ..... eta.	zenidan	zidan	nizun	zenidazun
75	Zuk erositakoareki n dago nirea.	Erosi duzula horrekin dago nirea.	Erosi duzunakin dago nirea.	Erosi horrekin duzuna dago nirea.	Erosi duzunarekin dago nirea.
76	Nik zuk ..... txikito edan dut.	beste	bezain	bezala	berdin
77	..... emango diote saria.	Lehenengoari iristen dela	Iristen den lehenengoari	Lehenengoa iristen delarik	Lehenengoa iristen da horri
78	Hura ..... den etxerik ez dut sekula ikusi.	hain polita	bezala polita	bezain polita	nola polita
79	Zu ariketak ondo ..... saiatzen zara?	egiten	egiteko	egitea	egin
80	Zein ..... bizi zara?	etxen	etxetan	etxeetan	etxean
81	Edozein ..... egin dezake hori.	gudarik	gudariek	gudariok	gudariak
82	..... joan ziren korrika.	Urera	Uratara	Uretara	Urara
83	Zein EZ dago ondo?	Zapatari	Idazkari	Kantalari	Zuzendari
84	Oso mesfidatia da, ez du ..... konfiantzarik?	inortaz	inorekin	inorengana	inorengan

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
85	Badaukazu konfiantzarik inorengan? ..... ez, baina lagunengan bai.	Edonorengandik	Edonorengan	Edonorengana	Edonon
86	Haurrak korrika ..... gurasoei ikastolatik eta "espaziozainak" ikusi nahi zutela esan zuten.	etorri zitzaizkien	etorri zitzaien	etorri zitzaizkizuen	etorri zitzaizuen
87	Nik ikasgaia azaldu nien, baina haiek dena	jakiten.	dakiten.	zekien.	zekiten.
88	Atzo guri Ameriketako senideak etorri	ziren.	zizkiguten.	zaizkigun.	zitzaizkigun.
89	"Cada vez que te veo me acuerdo del viaje a París."	Ikusten zaitudan bakoitzean, Pariserako bidaiaz oroitzen naiz.	Noiz eta ikusten zaitudanean, Pariserako bidaiaz oroitzen naiz.	Bakoitza ikusten zaitudala, Pariserako bidaiaz oroitzen naiz.	Ikusten zaitudala, Pariserako bidaiaz oroitzen naiz.
90	Nondik ..... zu atzo gaueko hamaiketan?	zetozten	zetorren	zentozten	zentozen
91	Zuk gurutzegrama hori daramazu haiek alaitzeko. Atzo	zeneraman.	eramaten zenuen.	zeramazun.	zeneramazun.
92	Aspaldiko lagun honek Baionaraino ekarri ..... gu, atzo.	genuen	zigun	genidan	gintuen
93	"Me llevaron con los ojos vendados".	zidaten.	nauten.	zitzaidan.	ninduten.

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	Begiak estalita eraman				
94	..... ere ez du alaitasunik ematen.	Aberatsa izatea	Aberatsa izateak	Aberatsa izaten	Aberatsa izateko
95	..... zaila gertatzen zaie lana aurkitzea.	Bilatzen ere badute,	Bilatu bada ere,	Bilatu arren,	Arren bilatu,
96	"Saber lo sabe", baina ez du adierazten.	Jakin badaki	Daki badaki	Dakiela badaki	Jakitea badaki
97	Zein da zuzena?	Ni Bilbokoa naiz; zu, ordea, Gernikakoa zara.	Ni Bilbokoa naiz; zu Gernikakoa berriz zara.	Ni Bilbokoa naiz; zu Gernikakoa zara berriz.	Bilbokoa naiz; ordea, zu Gernikakoa zara.
98	Zer egin diozue Txomini? "Muy enfadado" dago zuekin.	haserre zeharo	haserre oso	zeharo haserre	haserre guztiz
99	"Dejó de beber vino".	Ardoa edatetik utzi zitzaion.	Ardoa edateari utzi zion.	Ardoa edaten utzi zion.	Ardoa edatea utzi zuen.
100	"Le obligaron a pagar la multa".	Isuna ordaintzera behartu zuten.	Isuna ordaintzeko behartu zuten.	Isuna ordaintzen behartu zuten.	Isuna ordaintzea behartu zuten.
101	..... porrot egin zuen.	Ikastolen aldeko manifestazioak	Ikastolen alde manifestazioak	Manifestazioak ikastolen alde	Manifestazioa ikastolen aldekoa
102	Gurasoek egia osoa esatera behartu zuten mutila.	Gurasoek egia osoa esanarazi zioten mutilari.	Gurasoek egia osoa esan egin zioten mutilari.	Gurasoek egia osoa esan zioten mutilari.	Gurasoek egia osoa esan egin zuten mutila.
103	..... hori eta ikusiko duzu zer erantzuten dizun.	Esaiozu	Esan diozu	Esan zaio	Esan duzun
104	"La carretera hasta Ondarroa" biragune asko du.	Ondarroaraino errepideak	Ondarroaraino ko errepideak	Errepideak Ondarroaraino	Errepideak Ondarroara arte
105	Mezatan "bildu zen	bilduriko dirua	dirua bildu zela	bildutakoa dirua	diru bildutakoa

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	dirua", txiroentzat da.				
106	..... apurtu duzu.	Niregandik	Nigatik	Niregana	Nitaz
107	"Zuk egiten duzun bezala" egiten badugu guk, pikutara goaz.	Zuk egiten duzun moduan	Nola zuk egin duzu	Zuk egin duzun nola	Nola zuk egin duzun
108	Guk haiei postala idatziko	baliegu, ...	bagenie, ...	baligute, ...	bagenieke, ...
109	..., haiek guri beste gutun bat bidaliko	zigukete.	genieke.	ligute.	ligukete.
110	Haiek etxe garesti horietan biziko	balira	balute	balituzte	balirake
111	..., gela dotore eta handiak edukiko	lituzkete.	lukete.	lituzteke.	lirakete.
112	"Ayer todos lo llevaban menos tú".	Atzo denek zeramaten, zuk izan ezik.	Atzo denek izan ezik, zuk ez zeneraman.	Atzo denek zuk gutxiago zeramaten.	Atzo denek zeramaten, zuk izan gutxiago.
113	"Estando Caperucita Roja en el bosque" otsoa agertu zen.	Txano Gorritxo basoan zegoenetik	Txano Gorritxo basoan zegoenez	Txano Gorritxo basoan zegoenik	Txano Gorritxo basoan zegoela
114	Errealak ..... oso triste dago.	galdu zeneztik,	galdutzetik,	galdu zuenetik,	galdu zuenera arte,
115	Hau da lekua jaun-andreok, ... hil zuten Canovas.	hemendik bertan	hementxe	hor bertatik	han
116	Bakoitzari luma bat eman diot.	Lumana bat eman diet.	Luma bana eman diet.	Luma bati eman diet.	Lumak batari eman diet.
117	"Uno junto al otro" eseri	Elkarren ondoan	Elkar	Elkarna	Elkarrengana

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	ziren.				
118	"Mañana nos veremos".	Bihar ikusiko gara.	Bihar ikusiko dugu elkar.	Bihar ikusiko diogu elkarri.	Bihar ikusiko diogu.
119	Niri loteria ..... bidaia egingo nuke.	irtengo balit,	irtengo badit,	irtengo balio,	irtengo balitzait,
120	Zuk esaten duzun plan hori ongi irudituko ..... Koldori datorren iganderako.	zioke	litzaidake	lioke	litzaioke
121	Gustatu izan balitzaizue, erosi egingo	zenukete.	zenuketen.	zenuketeen.	zenuketeke.
122	Mikel ez da etorri, baina gurasoei egia esan	balie.	bazie.	baliote.	baziote.
123	Baimena izan balu, gurekin batera	etorriko litzateke.	etorriko litzatekeen.	etorriko zatekeen.	etorri litzateke.
124	Guk ..... eguzkia betiko ezkuta daiteke.	baldin abestu,	abestu baldin,	abestu baldin ezkero,	abestuz gero,
125	Trena abiatu ..... txartela puskatuta geneukan.	zenerakoan,	zela,	zenerako,	zenetikoan,
126	Dirua utziko dizut, baina gero	bueltatzekotan.	bueltatuz.	bueltatzerakoan.	baldin eta bueltatu.
127	Zarata handia ..... ondorioz gor gelditu gara.	egitearen	egiteko	eginez	egiteaz
128	Hori "haciéndolo" ikasten da.	eginez	eginerako	egiten	egindakoan
129	Ondoko dendan "robando" ikusi zuten.	lapurtuz	lapurtzen	lapurtzerako	lapurtzeko
130	Miren, bidal	dizkiozu	biezazkizu	iezazkiozu	zenizkiozu

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	..... Anderri eskutitz hauek.				
131	Jaietatik auto ..... etorri ginen.	banan	banaka	banatan	banan-banan
132	Gaztetan lagunei bisitatzera urtero joaten	nintzaien.	zitzaizkidan.	nintzaizkien.	natzaie.
133	Herenegun ez ..... (zuek gu) ikusi herriko plazan?	gintuzten	genituzten	gaituzue	gintuzuen
134	Gaur egun erretzaileok ez daukagu non-nahi (erre)	erretzea.	erretzerik.	erretzen.	erretzeari.
135	Negualdian etxea berotzeko ..... aritzen dira inguruko auzokoak.	erreketan	zamaketan	egurketan	olgetan
136	Garaiz bukatuz gero jaialdira joan ..... (zuek).	zaitezke	zaitezkete	zintezke	litezke
137	Ariketa hauek edonork egin ..... ondo.	daiteke	dezake	daitezke	ditzake
138	Ezin (nik zuei) ..... egia esan.	zeniezadakete	niezazuekete	diezazueket	diezazuket
139	Umeei ez ezik nagusiei ere gerta ..... horrelakorik.	dakieke	dakioke	dakiguke	dakizkieke
140	Eta hala ....., ez nizuke sinestuko.	izatea ere	baleike ere	izanda arren	balitz ere
141	Erromatarrak Euskal Herrira ere	antza	beharbada	orduko	omen

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	sartu ..... ziren.				
142	Medikuak ..... arabera urtebeteko bizitza baino ez dauka.	esandakoarekin	esanetan	esandakoan	esandakoaren
143	Festan ezagutu nuen pertsona ..... ikusi dut bulego horretan.	berdina	bera	berau	hauxe
144	..... ikasten da oinez.	Erorita	Eroririk	Eroriz	Erortzean
145	Desagertu bada ere ..... utzi nuen nik!	honaxe	horiexetan	horretantxe	hementxe
146	Ume nagusia lotsagabe galanta da, txikerra oso	lotsatua	lotsazalea	lotsakoia	lotsatia
147	Zuekin ez dago	ezer egiterik.	zerbait egiterik.	ezer egitekorik.	zer egitekorik.
148	Hau ez da batere erraza	egiteko	egiten	egiterik	egitekorik
149	Esan egia! ..... baten batek apurtu du kristala, ezta?	zutariko	ikasletariko	zuetariko	zuekiko
150	Ez diguzue batere lagundu, ..., oztopoak jarri dizkiguzue.	aitzitik	hala ere	baina	edonola ere
151	Joan ..... komunera, mesedez?	gaitezke	dezakegu	leikegu	genezake
152	Nahi izanez gero horiek ere ikas ..... euskara.	lezakete	lezake	litzakete	litezke
153	Hori soberan	diezaiekezue	zeniezaieke	zeniezaiokete	zeniezaiekete

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	balego, horiei eman ..... (zuek).				
154	Zorte apur batekin, gutxi erosten dutenei ere irten ..... loteria.	lekioke	lekieke	lekiguke	lekizkieke
155	Jon abilagoa izan balitz orain ez ..... bakarrik egongo (Jon).	zatekeen	ziratekeen	litzatekeen	liteke
156	Lagunekin ondo konpondu izan bazinete, haiekin eramango ..... (zuek).	zaituztete	zintuzteke	zintuztekete	zintuzteketen
157	Ezkonberritan zer egin genezakeen gurasoek umeak .....?	zaindu izan ez balizkigute	zaindu ez bazizkiguten	zainduko ez balizkigute	zaindu ez baziguten
158	Egin ..... zure nahia zeruan bezala lurrean ere!	bedi	bitez	beza	dezala
159	Eman ..... aldean daukazun guztia!	iezaidazu	iezaguzu	iezaiguzu	biezaio
160	Jarrai ..... atzetik! osterantzean, alde egingo dio-eta.	bekio	dadila	bedi	bezate
161	Zein da egokiena?	Bizimodua % 5a merketu da.	Bizimodua % 5ean merketu da.	Bizimodua % 5 merketu da.	Bizimodua % 5an merketu da.
162	Zuk ez badakizu ere, ... jakingo du hori.	nork	nor edo nor	inork ez	baten batek
163	Zein EZ	Eskukada	Haurkuntza	Hauskorra	Esnekia

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	dago ondo?				
164	Bihar arratsaldean deituko dizuet gurekin batera etor	zintezke.	zaitezte.	zaitezten.	zaitez.
165	Eutsi diru honi, zuk nahi duzuna eros	dezan.	dezaten.	dezadan.	dezazun.
166	Galduta nago; esan ..... zein den nire logela?	zeniezadake	didazu	dezakezu	iezadazu
167	Jesukristo hirugarren egunean ..... omen zen.	birlandatu	berrikusi	birsortu	berpiztu
168	Nola idazten da hizkiz '1639'?	Mila eta seiehun eta hogeita hemeretzi.	Mila seirehun eta hogeitahemeretzi.	Mila seiehun eta hogeita hemeretzi.	Mila seiehun eta hogei eta hemeretzi.
169	..... emango zenidake, mesedez?	Basokada bat esne beroa	Basokada bat esnea bero	Basokada esne bero	Basokada bat esne bero
170	Gaur eguraldi ona dugu, baina ....., biharko euria dakar.	dirudienez	antza dela	badirudi	nahitaez
171	Lana Kepak ....., eskerrak berari eman nizkion .	egin zidanalakoan	egin ditelakoan	egin zuelakoan	egin duelakoan
172	Venezuelan ..... izugarrizko hondamendia eragin dute.	haize boladak	euriteak	euriteek	eguraldiak
173	Saia zaitez ariketa hau ..... egiten.	ahalik eta hobetoen	ahalik eta ondoena	ahalik eta hobekien	ahalik eta ondoen
174	Nolakoa da	Irakurterraza	Irakurrerraza	Irakurri erraza	Irakurtierraza

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	testu hau?				
175	Ez nahastu bazterrak, ... egingo dugu eta.	hemendik bertan	gaur bertan	gaurxe bertan	gaur bertanxe
176	Gaur egungo arropek ez ..... ezer.	irauten dute	dirau	dihardute	diraute
177	Jesukristo mundura ..... !	baletor	bazetorren	balego	balebil
178	Bera ez ..... inor entzuteko prest.	leudeke	legoke	letorke	legokez
179	Gaur ez da etorri eta bihar ere ez da etorriko, ...	nonbait	antza dela	badirudi	litekeena da
180	Txiroak ..... eta ezjakinak ....., denok biziko ginateke hobeki.	balu/balitz	balekar/balu	baleki/baleuka	baleuka/baleki
181	Azken hamar urteotan egunero etorri ..... da lanera, baina gaur huts egin du.	nonbait	antza	liteke	ohi
182	Idazki hori ..... da.	ulergaitza	ulerrezina	ulerkaitza	ulertu gaitza
183	Baldintza hauek erabat ..... dira.	onartezinak	onarrezinak	onartuezinak	onar ezinak
184	Txapel horrekin deabruak ..... (zuek).	dirudite	dirudizu	iruditzen zarete	dirudizue
185	Egunkariak ..... EUROak behera egingo du.	dakarrenaren arabera	dakarrenaren arauera	segun eta dakarren arabera	dakartenaren arabera
186	..... ,	Badirudi	Badirudiela	Antza denez	Dirudiela

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	hurrengo hilabetera arte oporretan daude.				
187	Zenbat eta ....., orduan eta ..... duzu koxean sartzeko.	gizenagoa/zailago	gizenagoa/zailagoa	gizenago/zailago	gizenago/zailagoa
188	..... ederra egin du ..... gaztearekin.	Bidaia/anaia	Bidai/anaia	Bidaia/anai	Bidai/anai
189	Etorriko ..... ustean egon gara denok, baina ez da etorri.	dela	delaren	denaren	delako
190	Zuk ..... irratian entzun dut.	aipatunikoa	aipaturikoa	aipatuakoa	aipatu
191	Ainhoak ..... lan egiten du euskaltegian.	irakasle bezala	irakasle	irakasle moduan	irakasle gisa
192	Mutil hori adurra ..... gelditzen da neskei begira.	dariela	darizkiola	dariola	darizkiela
193	Ordenagailuz idazten ..... argia joan zen.	ari nintzela	nengoela eta	bitartean	nengoenez
194	Testu luzeak idazteko ordenagailua ..... beti.	darabildan	narabil	darabilt	dautza
195	Ainhoa aspertuta dago, ...	ni aspertuta nago ere bai.	ni aspertuta nago baita.	ni aspertuta ere nago.	ni ere aspertuta nago.
196	Peruri deitu nion gurekin hondartzara etor	nendin.	dadila.	zedin.	zedila.
197	Gizonak badu mundu honetan makina bat	eginkizun.	egikera.	eginkortasun.	egintza.

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
198	Guraso, anaia eta ..... artean garraiatu ditugu etxeko altzariak.	laurok	lauron	laurak	lauren
199	Lagunei deitu genien etxera etor	zekigun.	zekiguten.	zekizkigun.	zekizkiguten.
200	Eskerrak sasoiz heldu ....., bestela ez dakigu zer egingo genukeen!	zineten	zinetela	zinetelari	zineteni
201	Dirua utzi nizun handik aste betera itzul	zeniezadaten.	zeniezadatela.	zeniezadan.	niezazun.
202	Mozkor-mozkor eginda etorri zen, ...	dirudiela.	antzaz.	baliteke.	itxuraz.
203	Agindutako guztia bete ..... nahi zuten, baina ez zenien jaramon handirik egin.	zenezan	dezazun	zenezaten	zezaten
204	....., bilera amaitutzat jo dugu eguerdiko ordu bata denean.	Azkenean	Horienbestez	Honenbestez	Harrezkero
205	Zein da zuzena?	monetal politika	aho-literatura	unibertsitateko-esparrua	biologi-azterketa
206	Erdi lo nengoen eta ..... harrapatu ninduten.	ustekabean	ustegabean	uste gabean	uste kabean
207	Azken ordura arte ez dugu jakingo joango	garenen ala ez.	garenentz.	garela.	garenik.
208	.....	Zeuek	Zeuk	Zuk	Zuok

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	ikasleok, zer uste duzue dela bizimodua?				
209	Zeinek adierazten du ekintza?	Bultzada	Entzungailua	Pilotazalea	Txistularia
210	Ainhoak eta biok hanka sartu dugu, baina ikasleek ..... sartu dute, eta ez da inor ezertaz konturatu.	baita ere	ere bai	baita	ere
211	Subjuntiboa azaldu nizuen zuen nahi edo helburuak zuzen adieraz	zenitzan.	ditzazun.	zenitzaten.	zintezten.
212	Segituan deitu nien suhiltzaileei albait arinen etor	daitezen.	zitezkeen.	zitezen.	zintezten.
213	Zuri gertatutakoa guri ere gerta ..... nahi zenuen, ala?	zekiguten	dakigun	dakiguke	zekigun
214	Dirua eman nien musika irakasleari ordain	ziezaion.	ziezaioten.	zekion.	dakion.
215	Edonon egon ....., baina azkenean herrian bertan aurkitu genuen.	zitekeen	zitezkeen	ziratekeen	zintezkeen
216	Han egon izan bagina ere zer egin	genezake?	gintezkeen?	genitzakeen?	genezakeen?
217	Etorri ..... lanaren erdia	bitartean	zarenerako	zarenetik	arteko

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	amaituta nuen.				
218	..... interesgarria izan dela Urdaibaiko hegaztiak aztertzea!	Zeinen	Zein	Horra	Bai
219	Astiro azaldu izan banizu hurrengoei argi eta garbi azal	zeniezaiekeen.	zeniezaieke.	zeniezaieketen.	zeniezaiekete.
220	Hemen gauden Ainhoa, hiru ikasle eta ..... antolatuko ditugu datorren asteko jardunaldiak.	bostak artean	bosten artean	bostaren artean	boston artean
221	"Aiton-amonek makina bat gozoki erosten diete umeei". Inpertsonal bihurtu.	Makina bat gozoki erosten zaie umeei.	Makina bat gozoki erosten zaio umeei.	Makina bat gozoki erosten zaizkio umeei.	Makina bat gozoki erosten die umeei.
222	Neu ere joan ....., baina zeu zindoazela ikustean, atzera egin nuen.	nintzatekeen	nintzateke	ninteke	nintekeen
223	Duela ordu bete buka ..... lan hau, baina hemen gabiltza oraindino, jo eta ke.	genezake	gintezke	genezakeen	gintezkeen
224	Egia esan ....., argi eta garbi; baina	zeniezaieten	zeniezaieketen	zeniezadakeen	zeniezaiokeen

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	ohi bezala, gezurretan aritu zineten.				
225	Zuk ez dakit, baina nik, ..., oporrak hartzeko gogo izugarria daukat.	behinik-behin	gutxienez	behinik behin	behinipein
226	Dena dela, lanez gainezka gabiltzala-eta, ez ..... dugu jai egunik hartuko.	ote	al	ahal	bide
227	Lagunei deitu nien Aste Santuko oporretan guri etor	zekizkigukeen.	zekigukeen.	zekizkigukeelako.	zekizkigukeela.
228	Oporretako bidaia ..... diru kopurua baino, geure gustuak izan ditugu kontuan.	aukeratutakoan	aukeratu orduko	aukeratuz gero	aukeratzerakoan
229	Hauetako zein da aurrizkidun hitza?	Prakabarrenak	Zuzendariordea	Garbigailua	Basabelarra
230	Zer nolakoa duzu ..... hartu-emana?	amaginarrebarenganako	amaginarrebakiko	amaginarrebarekiko	amaginarrebarekin
231	..... errespetua ezinbestekoa da elkarbizitzarako.	Etxekoenkiko	Etxekoarenkiko	Etxekoarekin	Etxekoekiko
232	..... kafeari, azukrea neuk bota diot-eta.	Eragiozu	Eragin ezazu	Eragin zaitez	Eragin diozu
233	Bertan behera utziko dugu	besterik	ordea	osterantzean	ostantzean

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	gaurko lana; ....., baten bat akabatu egin behar dugu-eta!				
234	Lehengo egunean bazkaldu ..... jatetxea ez zen batere merkea.	genuenaren	genuelako	genuen	genueneko
235	" Udaberria gainean dugula ...". Zein da baliokidea?	Udaberria heltzear dagoela ...	Udaberria erortzeko zorian dagoela ...	Udaberria atarian dagoela ...	Udaberria heltzearen zorian dagoela ...
236	Zein EZ da zuzena?	Telefono-zenbakia	Besamotza	Eliz-gizona	Lotsabakoa
237	Ez dugu jakin zein zapatila hartu kirola egiteko; azkenean ..... hartu ditugu.	hauek berauek	beroriek	hauexek	horiexek
238	Jaiotza-tasa, ... oso baxua ....., igotzeko ahaleginetan dabil Jaurlaritza.	non/den	zein/baita	zeinen/baita	zeinetan/den
239	Berandutxo zabiltza, ezta? Zein EZ da baliokidea?	Berandu samar zabiltza, ezta?	Nahiko berandu zabiltza, ezta?	Berandu antzean zabiltza, ezta?	Pixka bat berandu zabiltza, ezta?
240	" Ondotxo daki horrek zer darabilen esku artean". Zein EZ da horren kidea?	Ondo baino hobeto daki horrek ...	Oso ondo daki horrek ...	Ondo bezain ondo daki horrek ...	Ondo daki, jakin ere, horrek ...
241	Ipuin hori ..... da.	ikasleagatik asmatua	ikasleak berak asmatua	ikasleagatik asmatuta	ikasleak berak asmatutak
242	"Entzun duzu Mozambikekoa?" Aukeratu	Halakorik!	Bai, arrano-pola!	Hori entzun behar genuen!	Bai entzun dudala!

ID	Enunciado	Opción1	Opción2	Opción3	Opción4
	erantzun zuzena.				
243	Aukeratu zuzen EZ dagoena.	Zaude lasai, jakinaraziko zaigu-eta!	Lanean jarraituko baduzu denetik egin beharko duzu.	Gogoan dut jaio zineneko eguna.	Ematen didazu txartel bat, mesedez?
244	Autoarekin ibiltzen ginen moduan ibilita, guri ere gerta ..... istripuren bat.	dakiguke	lekiguke	zitzaigun	zekigukeen
245	Zein EZ da zuzena?	Etorri izan balitz, entzungo zituen entzun beharrekoak.	Udaro Asturias aldera joateko ohitura dugu, baina aurten Mediterraneo aldera jo dugu.	Gurekin bazentoz, gurean gera zintezke.	Badirudi elkarrekin bizitzea ez denik nahikoa.
246	Zein EZ da zuzena?	Kantatzearekin utzi dut.	Kantatzen utzi diet.	Kantatzen utzi ditut.	Kantatzeari utzi diot.
247	Ezin dugu ahaztu badagoela	kontraesan.	kontraesanik.	kontraesanak.	kontra esanik.
248	Egungo ezkontzak, ..., arranditsuak dira.	antzinakoak konparatuz	antzinakoak konparatuta	antzinakoen aldean	antzinakoez konparatuta
249	"Joera honek, gure ustez, ..." Zein EZ da egokia?	bakarrik alde onak dauzka.	alde onak baino ez dauzka.	alde onak besterik ez dauzka.	alde onak baizik ez dauzka.
250	Zein da zuzena?	Gero eta jende gehiagok erretzeari utzi nahi dio.	Beti egin izan dut honela.	Urte askotan gauza berbera entzuten daramagu.	Horregatik erretzea leku publiko askotan debekatuta dago.
251	Zein da zuzena?	Beste egunean	Pasadan egunean	Pasa dan egunean	Lehengo egunean
252	Zeinetan EZ dago "bere/nire" posesiboa soberan?	Neskek bere ilea zaintzen dute.	Nire amak laguntzen dit arropa erosten.	Gerrerok bere hanka sartu eta baloia harrapatu du.	Nori berea Jainkoaren legea.

ANEXO 2 Cuestionario de la CE

En la prueba de campo PE1 los 252 ítems del banco se distribuyen en 8 cuestionarios. Cada cuestionario contiene 42 ítems, de los cuales los 12 primeros son ítems de anclaje y el resto específicos del cuestionario.

La estructura de los cuestionarios está formada por *una introducción*, por una recogida *datos personales*, *los ítems a valorar* y *las aportaciones propias*. Del apartado de ítems a valorar se han eliminado la mayoría de los ítems, ya que estos están disponibles en el anexo "Banco de Ítems Original".

La **introducción** presenta el objetivo de trabajo y agradece la voluntaria participación en el mismo. A continuación, se presenta las instrucciones de cumplimentación del cuestionario.

La recogida de **datos personales** del participante pretender recoger datos estadísticos del sujeto que cumplimenta el cuestionario como su edad, sexo, titulación superior, titulación lingüística y experiencia laboral en el área. Estos datos sirven para agrupar los resultados y deducir resultados estadísticos.

Posteriormente se presentan los **ítems a valorar**. Por cada uno de los ítems se requiere la determinación de la respuesta correcta del ítem en sí, la destreza que trabaja el ítem y el nivel de dificultad del ítem, ya que son los rasgo en base a los cuales se organizan los ítems en el sistema Boga-Hezinet.

Finalmente se solicitan **aportaciones propias** al participante. Se trata de una pregunta abierta para recoger comentarios y/o sugerencias.

Seguidamente se muestra íntegramente el cuestionario número 8 de la PE1.

Universidad del País Vasco
Euskal Herriko Unibertsitatea

LENGOAIA ETA SISTEMA INFORMATIKOAK SAILA
DEPARTAMENTO DE LENGUAJES Y SISTEMAS INFORMATICOS

Adituen eskutik

Euskarako item banku baten kalibraziorako

8. galdesorta

Nori zuzendua?

Euskarazko irakasleei

Euskal filologoei eta

IVAP PL4 lortu berria dutenei

2004ko Otsaila

ESKERTZA PARTEHARTZAILEEI

Adiskide agurgarria:

Gu UPV/EHUko GHyM ikerketa taldeko partaideak gara, eta,

ESHAPE: Evaluación de Sistemas Hipermedia Adaptativos Para la Educación

proiektuan, helduei euskara irakasteko modulu pedagogiko bat dugu aplikazio baten barnean item edo ariketa banku batez osaturik dagoen. Gure helburua item hauek adituen bidez eta automatikoki kalibratzea da, ondoren kalibrazio biak konparatuko ditugu eta emaitzak ekoitziko. Esperimentua ondo ateratzen bazaigu, item bidez osatutako bankuen kalibrazio prozesua automatizatzea espero dugu.

Baina kalibrazioa egin aurretik, zugana jotzen dugu bankuko item bakoitzaren zuzentasuna, euskara aldetik, erabakitzen lagun gaitzazun. Oinarria okerra bada, ondoren egingo diren ikerketa lanen ondorioak baliogabeak izango baitira.

Hori dela eta, zure laguntza desinteresatuaren eske gatozkizu, eta horretarako aski izango da ondorengo galdesortari erantzutea.

Zalantzarik izango bazenu edota taldekideei kontsultarik egin nahiko bazenie, 943.01.50.43 edo 946.01.44.68 telefonora dei eginez edota jiparsar@si.ehu.es edo javilo@si.ehu.es posta elektroniko helbidera idatziz egin dezakezu.

Amaitzeko, jakinarazi nahi dizugu izengabetuaren berme osoa duzula. Jasotako informazioa globalki esplotatuko da eta banakako emaitzak ez dira emango, ez eta galdesortaren erantzuleen identifikazioa zabalduko ere. Era berean, galdeketaren emaitza globalak behin esplotazio eta analisi fasea amaitu ondoren, nahi duten partehartzaileen esku egongo da.

Zure partaidetza aldez aurretik eskertuz.

Rosa Arruabarrena
Unibertsitate Eskolako Irakasle Titularra

Javier López Cuadrado
Arduraldi Osoko Irakasle Laguna

Lengoaiak eta Sistema Informatikoak saila.
GHyM ikerketa taldea http://ji.ehu.es/hyper/
Euskal Herriko Unibertsitatea/ Universidad del País Vasco

GALDESORTAREN EDUKIA

Galdesorta honek hiru atal ditu: bata erantzulearen ezaugarri pertsonalak jasotzeko, bestea item bankuaren kalibrazioa jasotzeko eta azkenekoa iruzkinak emateko. Itemak edozein bezeroren euskarazko jakite maila zehazteko erabiliko diren ariketa-galderak dira. Eta funtsean, aditua zaren aldetik, item horien kalibrazioa egitea da eskatzen dizuguna galdesorta honen bidez.

GALDESORTA OSATZEKO BETE BEHARREKO AGINDUAK

- **Ezaugarri pertsonalak (1 or.)**

1. atala osatzeko norberaren egoerari dagokion aukera hauta ezazu emandakoen artean. Era berean, tarte irekietan eskatzen zaizun informazioa idatz ezazu.

Bestalde, berriz ere zin egiten dizugu ematen dizkiguzun datuak konfidentzialtasun osoz erabiliko direla balio estatistikoak sortzeko.

- **Bankuko itemen kalibrazioa (3 or.)**

2. atala 42 ariketaz osaturik dago. Ariketa-galdera bakoitzarekin tratamendu berdina egitea eskatzen zaizu, hots, bakoitzeko zure ustezko 3 erantzun egokienak ematea. Horretarako galdera bakoitzeko egin beharrekoak honako hauek dira:

- Ariketa-galderari erantzun zuzena eman proposatutako lau erantzunetatik bakarra aukeratuz.

- Ariketa horrek euskarako ikaslearen zein trebetasun lantzeko erabilgarria den erabaki ezazu. Horretarako trebetasun zerrenda bat eskaintzen zaizu eta bertatik bakarra aukeratu beharko duzu. Ahal duzun neurrian eskainitako trebetasun arloen arteko bat aukeratu; besterik balitz, idatzi zure ustez landutako arloa.

- Ariketaren zailtasun maila zein den erabakitzea ere eskatzen zaizu. Horretarako zirkulu bakar batez ingura ezazu maila 1 eta 12 arteko balio eskalan (maila hauek HABEko curriculum zaharrarekin bat datozelarik).

Euskarako ikasle berriak ikasi beharrekoa 1 2 3 4 5 6 7 8 9 10 11 12 Ega mailako ikasleak ikasi beharrekoa

- **Iruzkinak eta bakoitzak lekarkeena (25 or.)**

3 atala hautazkoa da. Nahi izanez gero, galdesortaren amaieran hutsune bat erreserbatu dizugu bertan galdesorta honi buruz duzun iritzi librea eman ahal izateko.

KALIBRAZIO LANERAKO LAGUNGARRIAK

Kalibrazio lanean lagungarri suertatuko zaizkizulakoan ematen dizkizugu:

- bi itemen kalibrazioak eginak, erantzun-adibidetzat har ditzazun, eta
- euskarako gaitasun agiri desberdinen arteko baliokidetza taula.

Baliokidetza taula duen orria koadernotxotik aska dezakezu eskura izan dezazun ariketa-galderak erantzuteko garaian.

- **Bi itemen kalibrazioen adibideak dituzu hemen**:

1. *1, 2, 3*

   - ❑ **One, two, three.**
   - ❑ **Eins, zwei, drei.**
   - ☒ **Bat, bi, hiru.**
   - ❑ **Uno, due, tre.**

   Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	☒ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

   Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

   Ikasle berriak — EGA mailako ikasleak
   1 2 3 (4) 5 6 7 8 9 10 11 12
   ikasi beharrekoa — ikasi beharrekoa

2. *Norbait modu onez eta losintxaz ari zaizunean.* Aukeratu erantzun egokia

   - ❑ **Ez niri adarrik jo.**
   - ❑ **Zoaz antzarrak ferratzera.**
   - ❑ **Intxaurrak urrunetik hamalau, hurbildu eta lau.**
   - ☒ **Ez hasi nirekin zurikeriatan.**

   Zein trebetasuna lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	☒ **Besterik:** *Testuen ulermena*

   Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

   Ikasle berriak — EGA mailako ikasleak
   1 2 3 4 5 6 (7) 8 9 10 11 12
   ikasi beharrekoa — **ikasi beharrekoa**

1 ATALA: **EZAUGARRI PERTSONALAK**

1. *Adina* ("X" erabiliz aukeratu zure erantzuna):

❑ **29 edo gutxiago** ❑ **30-39** ❑ **40-49** ❑ **50-60** ❑ **60 edo gehiago**

2. *Sexua* ("X" erabiliz aukeratu zure erantzuna): ❑ **Gizonezkoa** ❑ **Emakumezkoa**

3. *Titulazio altuena eta hura lortutako urtea:*

- *Diplomatua/Teknikaria:*_____
- *Lizenziatua/Ingeniaria:*_____
- *Doktorea:* _____
- *Besterik:* _____

4. *Euskarako gaitasun agiri altuena:*

❑ **EGAI**	**VAP** ❑ **HE3** ❑ **HE4**	**IVAP-ekHezkuntzarako** ❑ **HE2**	**Besterik:** _____

5. *Ogibideak:*

	Iraupena (hilebetetan)	**Enpresa izena**	**Hiria**	**Azkena?** ("X" erabili)
Euskaltegiko irakaslea				
Euskal filologoa				
Euskarako Irakaslea				
Besterik. Zehaztu: _____ _____				

- **Hizkuntza Eskakizun (HE) eta agiri ofizialen baliokidetza taula:**

HABEren kurrikulua [1983] *urrats + aldia* *100-150 ordu urratsak.*	**HABE**ren kurrikulua [1989] maila	**HABE** kurrikulu berria [1999] *maila*	Hizkuntza Eskola Ofiziala (EOI)	**IVAP**-Euskal Administrazioko **Hizkuntza Eskakizunak** (*HE* - PL) profil linguistikoak	**IVAP**-H Euskal Administrazioko **Hezkuntzarako Hizkuntza Eskakizunak** (*HE* - PL) profil linguistikoak
1 urrats					
2 urrats					
3 urrats		1A			
4 urrats ≡1 aldia	A maila				
5 urrats					
6 urrats	B maila	1B		1 HE	
7 urrats					
8 urrats ≡2 aldia	C maila				
9 urrats		2	4º	2 HE	1
10 urrats					
11 urrats					
12 urrats≡3 aldia	EGA maila	3	5º ≡ Aptitud	3 HE	2
EGAtik gora		4		4 HE Teknikari Agiria	

Estamentuek, instituzio ofizialek edo hainbat erakundek ematen edo eskatzen dituzten Euskarako agiri, profil eta mailen arteko baliokidetzak

2 ATALA: BANKUKO ITEMEN KALIBRAZIOA

201

1. *Dirua utzi nizun handik aste betera itzul .....*

- ❑ **zeniezadaten.**
- ❑ **zeniezadatela.**
- ❑ **zeniezadan.**
- ❑ **niezazun.**

Zein trebetasun lantzeko erabiliko zenuke?

- ❑ **Aditzak**
- ❑ **Hiztegia**
- ❑ **Atzizkiak**
- ❑ **Deklinabidea**
- ❑ **Ortografia**
- ❑ **Idatzizko espresioa**
- ❑ **Sintaxia**
- ❑ **Lokailuak**
- ❑ **Besterik:**_____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

157

2. *Ezkonberritan zer egin genezakeen gurasoek umeak .....?*

- ❑ **zaindu izan ez balizkigute**
- ❑ **zaindu ez bazizkiguten**
- ❑ **zainduko ez balizkigute**
- ❑ **zaindu ez baziguten**

Zein trebetasun lantzeko erabiliko zenuke?

- ❑ **Aditzak**
- ❑ **Hiztegia**
- ❑ **Atzizkiak**
- ❑ **Deklinabidea**
- ❑ **Ortografia**
- ❑ **Idatzizko espresioa**
- ❑ **Sintaxia**
- ❑ **Lokailuak**
- ❑ **Besterik:**_____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

72

3. Zuk niri eskaini ..... liburuak eta zure lehengusuak aurpegi txarra jarri zuen.

- ❑ **zenizkidan**
- ❑ **zenizkidaten**
- ❑ **nizkizun**
- ❑ **zenizkioten**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

221

4. "Aiton-amonek makina bat gozoki erosten diete umeei". Inpertsonal bihurtu.

- ❑ **Makina bat gozoki erosten zaie umeei.**
- ❑ **Makina bat gozoki erosten zaio umeei.**
- ❑ **Makina bat gozoki erosten zaizkio umeei.**
- ❑ **Makina bat gozoki erosten die umeei.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

241

5. Ipuin hori ..... da.

- ❑ **ikasleagatik asmatua**
- ❑ **ikasleak berak asmatua**
- ❑ **ikasleagatik asmatuta**
- ❑ **ikasleak berak asmatutak**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

27

6. Alkandora ..... 2.000 pezeta balio du.

- ❑ **hau**
- ❑ **honek**
- ❑ **hauek**
- ❑ **hauk**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

92

7. Aspaldiko lagun honek Baionaraino ekarri ..... gu, atzo.

- ❑ **genuen**
- ❑ **zigun**
- ❑ **genidan**
- ❑ **gintuen**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

7

8. Zuek Donostian bizi al zarete?

- ❑ **Ez, gu ez bizi gara Donostian.**
- ❑ **Donostiakoak gaude gu.**
- ❑ **Ez, gu ez gara Donostian bizi.**
- ❑ **Ez, ez bizi gu gara Donostian.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

176

9. Gaur egungo arropek ez ..... ezer.

- ❑ **irauten dute**
- ❑ **dirau**
- ❑ **dihardute**
- ❑ **diraute**

Zein trebetasun lantzeko erabiliko zenuke?

- ❑ **Aditzak**
- ❑ **Hiztegia**
- ❑ **Atzizkiak**
- ❑ **Deklinabidea**
- ❑ **Ortografia**
- ❑ **Idatzizko espresioa**
- ❑ **Sintaxia**
- ❑ **Lokailuak**
- ❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

137

10. Ariketa hauek edonork egin ..... ondo.

- ❑ **daiteke**
- ❑ **dezake**
- ❑ **daitezke**
- ❑ **ditzake**

Zein trebetasun lantzeko erabiliko zenuke?

- ❑ **Aditzak**
- ❑ **Hiztegia**
- ❑ **Atzizkiak**
- ❑ **Deklinabidea**
- ❑ **Ortografia**
- ❑ **Idatzizko espresioa**
- ❑ **Sintaxia**
- ❑ **Lokailuak**
- ❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

47

11. ..... da trena hartzeko.

- ❑ **Beranduena**
- ❑ **Beranduegi**
- ❑ **Berandu gehiegi**
- ❑ **Berandu gehiago**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

112

12. "Ayer todos lo llevaban menos tú".

- ❑ **Atzo denek zeramaten, zuk izan ezik.**
- ❑ **Atzo denek izan ezik, zuk ez zeneraman.**
- ❑ **Atzo denek zuk gutxiago zeramaten.**
- ❑ **Atzo denek zeramaten, zuk izan gutxiago.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

51

13. Nire prakak ..... dira.

- ❑ **zureak bezalakoak**
- ❑ **zureak bezala**
- ❑ **zure bezain**
- ❑ **zure bezalakoak**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

227

14. Lagunei deitu nien Aste Santuko oporretan guri etor .....

- ❑ **zekizkigukeen.**
- ❑ **zekigukeen.**
- ❑ **zekizkigukeelako.**
- ❑ **zekizkigukeela.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

235

15." Udaberria gainean dugula ...". Zein da baliokidea?

- ❑ **Udaberria heltzear dagoela ...**
- ❑ **Udaberria erortzeko zorian dagoela ...**
- ❑ **Udaberria atarian dagoela ...**
- ❑ **Udaberria heltzearen zorian dagoela ...**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

185

16. Egunkariak ..... EUROak behera egingo du.

- ❑ **dakarrenaren arabera**
- ❑ **dakarrenaren arauera**
- ❑ **segun eta dakarren arabera**
- ❑ **dakartenaren arabera**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

17. Zein da zuzena?

- ❑ **Ni Bilbokoa naiz; zu, ordea, Gernikakoa zara.**
- ❑ **Ni Bilbokoa naiz; zu Gernikakoa berriz zara.**
- ❑ **Ni Bilbokoa naiz; zu Gernikakoa zara berriz.**
- ❑ **Bilbokoa naiz; ordea, zu Gernikakoa zara.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

9

18. Bizkaiko ..... goaz.

- ❑ **mendiara**
- ❑ **mendietara**
- ❑ **menditara**
- ❑ **mendia**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

101

19. ..... porrot egin zuen.

- ❑ **Ikastolen aldeko manifestazioak**
- ❑ **Ikastolen alde manifestazioak**
- ❑ **Manifestazioak ikastolen alde**
- ❑ **Manifestazioa ikastolen aldekoa**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

109

20. ..., haiek guri beste gutun bat bidaliko .....

- ❑ **zigukete.**
- ❑ **genieke.**
- ❑ **ligute.**
- ❑ **ligukete.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

151

21. Joan ..... komunera, mesedez?

- ❑ **gaitezke**
- ❑ **dezakegu**
- ❑ **leikegu**
- ❑ **genezake**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

25

22. Lehen egunkaria erosi dut.

- ❑ **Nork erosi du egunkaria?**
- ❑ **Noiz erosi duzu egunkaria?**
- ❑ **Lehen zer erosi duzu?**
- ❑ **Egunkaria erosi al duzu?**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

84

23. Oso mesfidatia da, ez du ..... konfiantzarik?

- ❑ **inortaz**
- ❑ **inorekin**
- ❑ **inorengana**
- ❑ **inorengan**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

252

24. Zeinetan EZ dago "bere/nire" posesiboa soberan?

- ❑ **Neskek bere ilea zaintzen dute.**
- ❑ **Nire amak laguntzen dit arropa erosten.**
- ❑ **Gerrerok bere hanka sartu eta baloia harrapatu du.**
- ❑ **Nori berea Jainkoaren legea.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

223

25. Duela ordu bete buka ..... lan hau, baina hemen gabiltza oraindino, jo eta ke.

- ❑ **genezake**
- ❑ **gintezke**
- ❑ **genezakeen**
- ❑ **gintezkeen**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

147

26. Zuekin ez dago .....

- ❑ **ezer egiterik**
- ❑ **zerbait egiterik**
- ❑ **ezer egitekorik**
- ❑ **zer egitekorik**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

46

27. Aukeratu zuzena.

- ❑ **Ez dago inor egon.**
- ❑ **Ez da inor ez egoten.**
- ❑ **Ez dago inor.**
- ❑ **Ez dagoen inor.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

59

28. Zuek nirekin ....., ez dizuet kontzerturako sarrerarik emango.

- ❑ **ez bazatoz**
- ❑ **etorriko ez bazarete**
- ❑ **ez bazara etorriko**
- ❑ **ez bazatozte**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

172

29. Venezuelan ..... izugarrizko hondamendia eragin dute.

- ❑ **haize boladak**
- ❑ **euriteak**
- ❑ **euriteek**
- ❑ **eguraldiak**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

134

30. Gaur egun erretzaileok ez daukagu non-nahi (erre) .....

- ❑ **erretzea.**
- ❑ **erretzerik.**
- ❑ **erretzen.**
- ❑ **erretzeari.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

197

31. Gizonak badu mundu honetan makina bat .....

- ❑ **eginkizun.**
- ❑ **egikera.**
- ❑ **eginkortasun.**
- ❑ **egintza.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

126

32. Dirua utziko dizut, baina gero .....

- ❑ **bueltatzekotan.**
- ❑ **bueltatuz.**
- ❑ **bueltatzerakoan.**
- ❑ **baldin eta bueltatu.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

34

33. ..... lan egiten du.

- ❑ **Astoek bezala**
- ❑ **Astoak bezalakoa**
- ❑ **Bezala astoek**
- ❑ **Astoa bezala**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

76

34. Nik zuk ..... txikito edan dut.

- ❑ **beste**
- ❑ **bezain**
- ❑ **bezala**
- ❑ **berdin**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

71

35. Ez da ..... guda atomikoa izatea.

- ❑ **hain zaila**
- ❑ **beste zaila**
- ❑ **bezain zaila**
- ❑ **zaila bezain**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

160

36. Jarrai ..... atzetik! osterantzean, alde egingo dio-eta.

- ❑ **bekio**
- ❑ **dadila**
- ❑ **bedi**
- ❑ **bezate**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

202

37. Mozkor-mozkor eginda etorri zen, ...

- ❑ **dirudiela.**
- ❑ **antzaz.**
- ❑ **baliteke.**
- ❑ **itxuraz.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

122

38. Mikel ez da etorri, baina gurasoei egia esan .....

- ❑ **balie**
- ❑ **bazie**
- ❑ **baliote**
- ❑ **baziote**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

177

39. Jesukristo mundura .....

- ❑ **baletor!**
- ❑ **bazetorren!**
- ❑ **balego!**
- ❑ **balebil!**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

210

40. Ainhoak eta biok hanka sartu dugu, baina ikasleek ..... sartu dute, eta ez da inor ezertaz konturatu.

- ❑ **baita ere**
- ❑ **ere bai**
- ❑ **baita**
- ❑ **ere**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

21

41. Telebista ikusten ari al zara?

- ❑ **Bai, telebista ikusten naiz.**
- ❑ **Bai, telebista naiz ikusten ari.**
- ❑ **Bai, telebista ikusten ari naiz.**
- ❑ **Bai, ikusten naiz telebista ari.**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

248

42. Egungo ezkontzak, ..., arranditsuak dira.

- ❑ **antzinakoak konparatuz**
- ❑ **antzinakoak konparatuta**
- ❑ **antzinakoen aldean**
- ❑ **antzinakoez konparatuta**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Lokailuak**	❑ **Besterik:** _____

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak EGA mailako ikasleak

1 2 3 4 5 6 7 8 9 10 11 12

ikasi beharrekoa ikasi beharrekoa

3 ATALA: **BAKOITZAK LEKARKEENA**

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Berriz ere GHyM ikerlariek eskerrak ematen dizkizugu zure partaidetzagatik.

ANEXO 3 Resultados de la calibración CE

La muestra contiene 3315 entradas sobre 192 ítems emitidas por un total de 111 expertos designándola por el triple (m=3315; n=192; e=111).

En la sección 1 se presenta la estimación de los parámetros *Dificultad* y *Destreza gramatical*. Para ello se emplean los estadísticos M.dif y M.est.

En la sección 2 se describen los análisis de fiabilidad realizados a los resultados de la estimación.

1 Frecuencias de los pronósticos otorgados por los expertos

La siguiente tabla recoge las frecuencias relativas de los pronósticos otorgadas por los expertos sobre la dificultad de los ítems. Se han coloreado los juicios considerados por el método estadístico M.dif para calcular las estimaciones, concretamente, en amarillo los del criterio M.dif-1 y en naranja los considerados por M.dif-2.

	Dificultad pronosticada												
Ítem	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**#Juicios**
1	12	2											14
2	8	5	2	1	1	1							18
3	11	2	3	1									17
4	8	3	1	2	1		1						16
5	9	4	2	1									16
6	9	2	2										13
7	36	12	9	4	2								63
8	6	5	2	3	1			1					18
9	3	7	4			1	1						16
11	6	4	1		1			1	1			1	15
12	4	3	2	2	2	1							14
13	3	5	3	2	1	2							16
14	1			6	7								14
15	11	4	1				1						17
17		1	4	4	3	3							15
18	6	3	1	1						1			12
19	5	6	2		1								14
20			1	6	4	3	1	1					16
21	6	6		2									14
22	3	4	4	3		1							15
23	5	5	1	2									13
26	4	5	1	1			1	1					13
27	27	14	9	4	2	1	1						58
28	9	3	2					1	1				16
29	3	8	3			2							16
31	3	4	6	4									17
32		3	2	6	2	1							14
33		7	4	3	1	1							16
34		1	2	3	3	2		1		1			13
35		1	2	4	4	1	2						14
36		5	8	3						1			17
37		1	3	5	2	2	1						14
38		3	6	1	4	2							16
39	8	4		1									13
40	2	1	5	1	4	4							17
41		5		5	2	2	1						15
42		3	3	4	3	2		1					16
43		3	2		5	4	1						15
44		3	3	1	2		1			2			12
45			8	8	1			1					18
46		4	5	5	1	1	1						17
47		9	27	16	4	3	1		1				61
48			8	5	1								14
49		1	4	2	5	1							13
52		1	10	2									13

	Dificultad pronosticada												
Ítem	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**#Juicios**
53		1	9	4	2			1					17
54		2	6	7	1	1							17
55		7	4	2	2		1						16
57	4	4	1	3	1								13
58		1	2	6	6		1						16
61		1	3	2	5	4	2						17
64		2	5	4	1	1		1					14
65			10	3	1	2					1		17
66	2	2	2	6	1	1	1						15
67			3	3	5	3	2						16
68		1	1	5	3	4						1	15
69				5	6		2		1				14
70			1	6	6	2		1					16
71				9	6								15
72			5	31	20	5	5						66
73				9	5	1	1						16
74			2	4	4	2	1						13
75			1	4	5	4		2					16
76		1		9	4	1							15
77			6	6	1	1							14
78			2	4	5	5	1						17
79	1		4	2	5	1	1						14
81	1		3	1	5	5	1						16
82		1	2	4	3	1				1			12
84				5	4	5	2	1					17
85				5	3	5	1						14
86			1	4	5	2	2	1	2				17
87			1	5	6	2	1						15
88			2	8	3	2							15
89				4	5	2	2			1			14
90			1	6	5	2		1	2				17
91				2	6	4	1	2	1				16
92			6	15	30	5	3	1	1				61
93				5	4	4	1						14
94				3	6	5	2						16
95				4	4	2		1	1				12
96				1	1	3	5	4	3				17
97			1	3	6	3	2	1					16
98		1	1	2	5	4	3						16
99				1	5	2	3	1	1				13
100				5	3	7	1	1					17
101				5	2	5	1	1	1	1			16
102						4	6	1	1			1	13
104			1		4	4	2	3	1	1			16
106		1	1	6	2	3	2					1	16
107				2	5	2	1	2	1				13
108					5	5	6						16
109					5	4	2	2			1		14
110				1	8	5							14
111					6	8				3			17
112			2	5	17	23	8	4	1			1	61
113			1		4	6	4						15
114			1	3	4	5	2		1				16
116				1	3	8	3	3					18
117				3	5	4	1	1		1			15
118		1		1	4	5	3		1				15
119					8	4	3						15

	Dificultad pronosticada												
Ítem	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**#Juicios**
120					6	2	3	2					13
121						6	4	4	1				15
122					6	1	4	2					13
123					3	4	3	2	2	1			15
124				2	3	5	2						12
125					4	5	2	2	2				15
126					1	1	5	3	4	1			15
127					2	3	2	5	1	1			14
128			1			6	6	2	1		1		17
129			3		6	4	1	1					15
130				2	1	5	5	1	1			1	16
132							2	7		3	1	1	14
133				2	8	5	1						16
137		11	12	12	11	9	2	4					61
138							3	7	4	2			16
139						2	2	2	5	2	3		16
140					4	2	3	3	1	1			14
141			1	1	3	5	3	2					15
142			1	1	4	7	2			1			16
146		1	1	4	4	2	2	1					15
147					1	4	4	4	2			1	16
149					3	1	4	2	2	1			13
150					1	2	2	4	6			1	16
152						2	5	2	2		2		13
153						1	3	2	3	3	3		15
154								2	1	4	5	3	15
155					2	5	4	2		1		1	15
156						5	5	4	1			1	16
158								1	3	2	2	5	13
159				1	1	2	7	4	1	1	1		18
160									1	3	4	7	15
162	1			2	2	4	7			1			17
164						2	3	6		1	1	1	14
165							4	4	2	1		1	12
166							2	4	3	5	2		16
167					1	3	7	1	1	2	1		16
171					1		11	4	1				17
172		2	2	5	3	3	1						16
173			1		1	2	6	3	2	1		1	17
174				2	3		4	3		1			13
175		1		2	6	3	1	5					18
177								4	2	3	4	2	15
179						4	4	6		2			16
180							1	2	5	4	2	3	17
181				4	2	2	6	2	2				18
182				1	6	4	3		1				15
183				1	2	1	6	2		1			13
184					3	2		5	2	1			13
185				1	2	6	4	1	1		1		16
186				4	2	3	2	1	1				13
187				2		5	5	3	2				17
189							3	4	4		4	1	16
190				7	1	3		2	2				15
192				1			2	3	3	3	2	1	15
193				2	3	4	4	1		2			16
194							2	4	2	3	2	2	15
195			1	2	3	3			1				10

	Dificultad pronosticada												
Ítem	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**#Juicios**
196						1		7	2	2	1	1	14
197				2	4	3	6		1				16
198			1	6	2	4	1					1	15
203							2	1	4	2	1	4	14
206		1		4	1	4	2					1	13
210	1	1		1	4	2	1	3					13
212						1		8	2	2	2	1	16
213							2		1	1	3	7	14
214						1	3	2	4	2		2	14
215						2	6	4	1	2			15
216						3	2	7	3	1		1	17
217				2	3	6	3	1		1		1	17
219						1	3	4	3	2	1	1	15
220				2	4	3	3			1			13
221		1	9	14	14	14	3	2	2	1	1	2	63
222						2	4	5	2	2			15
223					1		4	5	2	1	2		15
224							2	4	4	3	1	1	15
226						1	3	3	2	4	2		15
227										1	3	11	15
228				1	2	7	3	2					15
229				1		3	2	4	3	1	1		15
230						3	2	3	3	1	1	2	15
231					1		1	8		3	1	2	16
238								2	1	3	3	6	15
241		1		2	9	12	14	11	5	2		3	59
244								3	4	4	2	2	15
245		1		2	1	5	2	4				1	16
246				1	5	3	1	2				1	13
247		1		2	2	4	4	4		2			19
250						3	2	4		1	1		11
251		3	7	2			1	1					14
252					1	3	5	4			1	2	16

Tabla 1.- Frecuencias relativas de los pronósticos de los expertos sobre la dificultad de los ítems (m=3315, n=192, e=111)

Obsérvese que el estadístico únicamente emplea 2933 juicios, aunque el tamaño de la muestra sea m=3315.

2 Valores estimados

En la siguiente tabla se muestran los detalles de la calibración de ítems empleando juicios emitidos por expertos. La columna *Id ítem* corresponde al identificador del ítem; las columnas *moda* y *media* recogen la moda y el promedio de las dificultades de los ítems; las columnas *D*, σ y *IC95* corresponden a la dificultad estimada por M.dif, su correspondiente desviación estándar y su intervalo de confianza al 95%; las columnas *Destreza* y *%Moda* recogen la destreza gramatical estimada que trabaja el ítem y el porcentaje de expertos que coinciden con dicha estimación; y la columna *Detalle* recoge motivos de eliminación de los ítems o las causas por las que se consideran potencialmente erróneos.

	Dificultad					Destreza		
Item	**moda**	**media**	**M.dif**	**σ(M.dif)**	**IC95 (M.dif)**	**M.est**	**%Moda**	**Detalle**
1	1	1,1429	1,1429	0,3499	[0,9772, 1,3085]	Sintaxis	71,43%	Marcado: contenido
2	1	2,1667	1,75	0,9014	[1,3550, 2,1450]	Sintaxis	53,33%	
3	1	1,6471	1,6471	0,9666	[1,2377, 2,0564]	Declinación	75,00%	
4	1	2,3125	1,7857	1,0809	[1,2741, 2,2973]	Sintaxis	66,67%	
5	1	1,6875	1,6875	0,9164	[1,2859, 2,0891]	Sintaxis	54,55%	
6	1	1,4615	1,4615	0,7458	[1,0929, 1,8301]	Declinación	66,67%	
7	1	1,7937	1,6885	0,9503	[1,4852, 1,8918]	Sintaxis	83,33%	
8	1	2,6111	2,125	1,111	[1,6381, 2,6119]	Sintaxis	54,55%	
9	2	2,6250	2,0714	0,7035	[1,7385, 2,4044]	Declinación	100,00%	
10								Retirado: C.it-2(2º)
11	1	3,4000	1,5455	0,6556	[1,1873, 1,9036]	Declinación	50,00%	
12	1	2,8571	2,1818	1,1134	[1,5735, 2,7901]	Sintaxis	83,33%	
13	2	2,9375	2,3077	0,9911	[1,8179, 2,7975]	Declinación	87,50%	Marcado: contenido
14	5	4,2857	4,5385	0,4985	[4,2921, 4,7848]	Conectivas	53,85%	
15	1	1,7059	1,375	0,5995	[1,1123, 1,6377]	Declinación	81,82%	
16								Retirado: C.it-1
17	3	4,2000	4,3571	1,1089	[3,8323, 4,8820]	Sintaxis	54,55%	
18	1	2,4167	1,7273	0,9621	[1,2016, 2,2529]	Declinación	44,44%	
19	2	2,0000	1,7692	0,6966	[1,4250, 2,1135]	Verbos	66,67%	
20	4	5,0000	4,8	1,0456	[4,3246, 5,2754]	Sintaxis	80,00%	
21	1	1,8571	1,8571	0,9897	[1,3887, 2,3256]	Verbos	45,45%	
22	2	2,7333	2,5	1,0522	[2,0020, 2,9980]	Verbos	58,33%	
23	1	2,0000	2	1,0377	[1,4871, 2,5129]	Verbos	75,00%	
24								Retirado: C.it-2
25								Retirado: C.it-1
26	2	2,7692	1,9091	0,9	[1,4174, 2,4008]	Sintaxis	81,82%	
27	1	2,0862	1,8148	0,9637	[1,5940, 2,0357]	Declinación	94,00%	
28	1	2,3750	1,5	0,7319	[1,1536, 1,8464]	Verbos	100,00%	
29	2	2,5000	2	0,6547	[1,6901, 2,3099]	Declinación	93,33%	
30								Retirado: C.it-2
31	3	2,6471	2,6471	1,0256	[2,2127, 3,0814]	Declinación	92,31%	
32	4	3,7143	3,5385	1,0088	[3,0399, 4,0371]	Sintaxis	92,31%	
33	2	3,0625	2,8667	0,9568	[2,4316, 3,3017]	Verbos	92,86%	
34	4	5,0000	4,5	1,0247	[3,9060, 5,0940]	Sintaxis	100,00%	
35	4	4,5714	4,7692	1,2499	[4,1515, 5,3870]	Declinación	100,00%	
36	3	3,2941	2,875	0,696	[2,5700, 3,1800]	Sintaxis	85,71%	
37	4	4,2857	4,25	1,0104	[3,7262, 4,7738]	Declinación	80,00%	
38	3	3,7500	3,4286	1,1157	[2,9005, 3,9567]	Declinación	100,00%	
39	1	1,5385	1,5385	0,8427	[1,1220, 1,9549]	Sintaxis	90,00%	
40	3	3,9412	4,5	1,2392	[3,9134, 5,0866]	Declinación	87,50%	
41	2	3,9333	3,3333	1,1785	[2,7223, 3,9443]	Sintaxis	91,67%	
42	4	4,1250	3,5385	1,0824	[3,0035, 4,0734]	Verbos	92,31%	
43	5	4,5333	5	1,0445	[4,4294, 5,5706]	Declinación	100,00%	
44	2	4,6667	3,2222	1,1331	[2,5197, 3,9248]	Sintaxis	80,00%	
45	3	3,8333	3,5882	0,5999	[3,3342, 3,8423]	Sintaxis	94,44%	
46	3	3,5882	3,2	0,9092	[2,7866, 3,6134]	Sintaxis	57,14%	
47	3	3,5574	3,2679	0,8126	[3,0850, 3,4507]	Sintaxis	69,05%	
48	3	3,5000	3,5	0,6268	[3,2033, 3,7967]	Sintaxis	100,00%	
49	5	4,0769	4,0769	1,141	[3,5130, 4,6408]	Verbos	100,00%	
50								Retirado: C.it-2
51								Retirado: C.it-2
52	3	3,0769	3,0769	0,4742	[2,8426, 3,3113]	Sintaxis	91,67%	
53	3	3,7059	3,4375	0,7881	[3,0921, 3,7829]	Verbos	94,12%	
54	4	3,5882	3,4375	0,7881	[3,0921, 3,7829]	Sintaxis	64,29%	
55	2	3,1875	2,9333	1,0625	[2,4502, 3,4164]	Sintaxis	63,64%	

	Dificultad					Destreza		
Item	**moda**	**media**	**M.dif**	**σ(M.dif)**	**IC95 (M.dif)**	**M.est**	**%Moda**	**Detalle**
56								Retirado: C.it-2
57	1	2,4615	2,25	1,1637	[1,6467, 2,8533]	Verbos	84,62%	
58	4	4,3125	4,1333	0,8844	[3,7312, 4,5355]	Conectivas	80,00%	
59								Retirado: C.it-2
60								Retirado: C.it-2
61	5	4,8235	4,7143	1,0973	[4,1949, 5,2337]	Sufijos	78,57%	
62								Retirado: C.it-2
63								Retirado: C.it-2
64	3	3,8571	3,3333	0,8498	[2,8927, 3,7739]	Sintaxis	84,62%	
65	3	4,1176	3,6875	1,044	[3,2300, 4,1450]	Sintaxis	92,86%	
66	4	3,6000	3	1,1547	[2,4013, 3,5987]	Sintaxis	54,55%	
67	5	4,8750	4,5714	1,0498	[4,0745, 5,0683]	Declinación	100,00%	
68	4	5,0667	4,7692	0,973	[4,2883, 5,2501]	Declinación	100,00%	
69	5	5,2143	4,9231	0,997	[4,4303, 5,4159]	Declinación	91,67%	
70	4	4,8125	4,6	0,8	[4,2362, 4,9638]	Sintaxis	64,29%	
71	4	4,4000	4,4	0,4899	[4,1772, 4,6228]	Sintaxis	100,00%	
72	4	4,6061	4,6061	0,9982	[4,4008, 4,8114]	Verbos	100,00%	
73	4	4,6250	4,625	0,857	[4,2494, 5,0006]	Verbos	100,00%	
74	4	4,6923	4,5	0,9574	[4,0036, 4,9964]	Verbos	100,00%	
75	5	5,2500	4,8571	0,9147	[4,4242, 5,2901]	Sintaxis	93,33%	
76	4	4,2667	4,2667	0,8537	[3,8785, 4,6549]	Sintaxis	100,00%	
77	3	3,7857	3,7857	0,8601	[3,3786, 4,1928]	Sintaxis	91,67%	Marcado: contenido
78	5	4,9412	4,8125	1,0136	[4,3683, 5,2567]	Sintaxis	100,00%	
79	5	4,2143	4,25	1,0104	[3,7262, 4,7738]	Sintaxis	81,82%	
80								Retirado: C.it-2
81	5	4,7500	4,8571	1,1249	[4,3247, 5,3896]	Declinación	92,86%	
82	4	4,5833	4,3	0,9	[3,7783, 4,8217]	Declinación	100,00%	
83								Retirado: C.it-2
84	4	5,4118	5,25	1,0308	[4,7983, 5,7017]	Declinación	100,00%	
85	4	5,1429	5,1429	0,9897	[4,6744, 5,6113]	Declinación	100,00%	
86	5	5,6471	5,1538	1,0263	[4,6466, 5,6611]	Verbos	100,00%	
87	5	4,8000	4,8	0,9798	[4,3545, 5,2455]	Verbos	100,00%	
88	4	4,3333	4,3333	0,8692	[3,9381, 4,7286]	Verbos	100,00%	
89	5	5,5000	5,1538	1,0263	[4,6466, 5,6611]	Sintaxis	92,31%	
90	4	5,2941	4,5714	0,8207	[4,1830, 4,9599]	Verbos	100,00%	
91	5	5,8750	5,6667	1,1926	[5,1244, 6,2089]	Verbos	100,00%	
92	5	4,8525	4,6071	0,7946	[4,4283, 4,7860]	Verbos	98,31%	
93	4	5,0714	5,0714	0,961	[4,6166, 5,5263]	Verbos	91,67%	
94	5	5,3750	5,375	0,927	[4,9687, 5,7813]	Sintaxis	61,54%	
95	4	5,4167	4,8	0,7483	[4,3662, 5,2338]	Sintaxis	100,00%	
96	7	7,1176	7,4667	1,0242	[7,0010, 7,9323]	Exp. escrita	38,46%	
97	5	5,3125	5,2857	0,9583	[4,8321, 5,7393]	Conectivas	64,29%	
98	5	5,1875	5,5714	0,9794	[5,1079, 6,0350]	Vocabulario	30,00%	
99	5	6,0769	5,8333	1,1426	[5,2409, 6,4257]	Sintaxis	87,50%	
100	6	5,4118	5,25	0,9682	[4,8257, 5,6743]	Sintaxis	66,67%	
101	4	5,8750	5,1538	1,0263	[4,6466, 5,6611]	Declinación	61,54%	
102	7	7,3077	6,9167	0,862	[6,4698, 7,3636]	Verbos	50,00%	Marcado: contenido
103								Retirado: C.it-2
104	5	6,5000	6,3077	1,1358	[5,7464, 6,8690]	Declinación	92,31%	
105								Retirado: C.it-2
106	4	5,1875	5,0769	1,141	[4,5130, 5,6408]	Declinación	100,00%	
107	5	5,9231	5,6667	1,3123	[4,9863, 6,3471]	Sintaxis	90,00%	
108	7	6,0625	6,0625	0,8268	[5,7002, 6,4248]	Verbos	87,50%	
109	5	6,4286	6,0769	1,0714	[5,5474, 6,6065]	Verbos	100,00%	
110	5	5,2857	5,2857	0,589	[5,0069, 5,5645]	Verbos	100,00%	

	Dificultad					Destreza		
Item	**moda**	**media**	**M.dif**	**σ(M.dif)**	**IC95 (M.dif)**	**M.est**	**%Moda**	**Detalle**
111	6	6,3529	5,5714	0,4949	[5,3372, 5,8057]	Verbos	100,00%	
112	6	5,8689	5,6415	0,8488	[5,4452, 5,8379]	Sintaxis	82,69%	
113	6	5,8000	6	0,7559	[5,6422, 6,3578]	Sintaxis	92,31%	
114	6	5,5000	5,4286	0,9794	[4,9650, 5,8921]	Sintaxis	66,67%	
115								Retirado: C.it-2(2º)
116	6	6,2222	6,3529	0,9666	[5,9436, 6,7623]	Sintaxis	61,54%	
117	5	5,7333	5,2308	0,8904	[4,7907, 5,6709]	Declinación	50,00%	
118	6	5,7333	5,7692	0,8904	[5,3291, 6,2093]	Sintaxis	58,33%	
119	5	5,6667	5,6667	0,7888	[5,3080, 6,0253]	Verbos	100,00%	
120	5	6,0769	6,0769	1,141	[5,5130, 6,6408]	Verbos	100,00%	
121	6	7,0000	7	0,9661	[6,5607, 7,4393]	Verbos	100,00%	
122	5	6,1538	6,1538	1,1666	[5,5773, 6,7304]	Verbos	100,00%	
123	6	6,9333	6,3333	1,0274	[5,8007, 6,8660]	Verbos	100,00%	
124	6	5,5833	5,5833	0,9538	[5,0888, 6,0778]	Sintaxis	75,00%	
125	6	6,5333	6,1538	1,0263	[5,6466, 6,6611]	Sintaxis	92,86%	
126	7	7,7333	7,7692	0,973	[7,2883, 8,2501]	Sintaxis	78,57%	
127	8	7,2143	6,8333	1,1426	[6,2409, 7,4257]	Sintaxis	61,54%	
128	6	6,8824	6,8667	0,8844	[6,4645, 7,2688]	Sintaxis	71,43%	
129	5	5,2000	4,8462	1,0987	[4,3031, 5,3892]	Sintaxis	66,67%	
130	6	6,6875	6	1,0377	[5,4871, 6,5129]	Verbos	100,00%	
131								Retirado: C.it-2
132	8	8,7857	8,3333	1,0274	[7,8007, 8,8660]	Verbos	100,00%	
133	5	5,3125	5,3125	0,768	[4,9759, 5,6491]	Verbos	100,00%	
134								Retirado: C.it-2
135								Retirado: C.it-2
136								Retirado: C.it-2
137	3	4,2787	3,5	1,0984	[3,2273, 3,7727]	Verbos	100,00%	
138	8	8,3125	8,3125	0,9164	[7,9109, 8,7141]	Verbos	100,00%	
139	9	8,7500	9,5	1,0408	[8,9604, 10,0396]	Verbos	100,00%	
140	5	6,8571	6,4167	1,1873	[5,8011, 7,0322]	Sintaxis	83,33%	
141	6	5,9333	6,3077	0,9911	[5,8179, 6,7975]	Sintaxis	88,89%	
142	6	5,8125	5,7143	0,7954	[5,3378, 6,0908]	Sintaxis	81,82%	
143								Retirado: C.it-2
144								Retirado: C.it-2
145								Retirado: C.it-2
146	4	5,0000	5,1667	1,0672	[4,6134, 5,7200]	Sufijos	53,85%	Marcado: contenido
147	6	7,4375	7,2857	1,0302	[6,7981, 7,7733]	Sintaxis	76,92%	
148								Retirado: C.it-2
149	7	7,1538	6,5	1,118	[5,8519, 7,1481]	Declinación	91,67%	
150	9	8,0625	8	1,069	[7,4940, 8,5060]	Conectivas	80,00%	
151								Retirado: C.it-1
152	7	7,9231	7,3636	0,9791	[6,8287, 7,8986]	Verbos	100,00%	Marcado: contenido
153	7	8,8667	8,5455	1,1571	[7,9133, 9,1776]	Verbos	100,00%	
154	11	10,4000	10,7692	0,8904	[10,3291, 11,2093]	Verbos	100,00%	
155	6	7,0667	6,4615	0,9295	[6,0022, 6,9209]	Verbos	100,00%	
156	6	7,3750	7,0667	0,9286	[6,6445, 7,4889]	Verbos	100,00%	
157								Retirado: C.it-2
158	12	10,5385	10,75	1,2332	[10,1106, 11,3894]	Verbos	100,00%	
159	7	7,3333	7,1333	0,9568	[6,6983, 7,5684]	Verbos	100,00%	
160	12	11,1333	11,1333	0,9568	[10,6983, 11,5684]	Verbos	100,00%	
161								Retirado: C.it-2
162	7	6,0000	6,0667	1,0625	[5,5836, 6,5498]	Declinación	75,00%	
163								Retirado: C.it-2
164	8	8,1429	7,3636	0,7714	[6,9422, 7,7851]	Verbos	100,00%	
165	7	8,3333	8	0,9535	[7,4791, 8,5209]	Verbos	100,00%	

	Dificultad					Destreza		
Item	**moda**	**media**	**M.dif**	**σ(M.dif)**	**IC95 (M.dif)**	**M.est**	**%Moda**	**Detalle**
166	10	9,0625	9,0625	1,2484	[8,5154, 9,6096]	Verbos	100,00%	
167	7	7,5000	6,8462	0,9484	[6,3774, 7,3149]	Vocabulario	81,25%	
168								Retirado: C.it-2
169								Retirado: C.it-2
170								Retirado: C.it-2
171	7	7,2353	7,375	0,5995	[7,1123, 7,6377]	Sintaxis	78,57%	
172	4	4,3750	4,5385	1,0088	[4,0399, 5,0371]	Declinación	64,29%	
173	7	7,4118	7,3846	0,9231	[6,9284, 7,8408]	Sintaxis	87,50%	
174	7	6,5385	6,7	1,1874	[6,0117, 7,3883]	Sufijos	58,33%	
175	5	5,8333	6,3333	1,2996	[5,7424, 6,9242]	Exp. escrita	36,36%	
176								Retirado: C.it-2
177	8	9,8667	9,5385	1,2163	[8,9373, 10,1396]	Verbos	100,00%	
178								Retirado: C.it-2
179	8	7,5000	7,1429	0,833	[6,7486, 7,5371]	Conectivas	66,67%	
180	9	9,7647	10,2143	1,1451	[9,6723, 10,7563]	Verbos	83,33%	
181	7	6,3333	5,7143	1,2778	[5,1095, 6,3191]	Verbos	56,25%	
182	5	5,8667	5,6429	0,895	[5,2192, 6,0665]	Sufijos	100,00%	
183	7	6,7692	6,7273	0,9621	[6,2016, 7,2529]	Sufijos	63,64%	
184	8	7,3077	6,7	1,3454	[5,9202, 7,4798]	Verbos	100,00%	
185	6	6,6250	6,1429	0,9897	[5,6744, 6,6113]	Sintaxis	72,73%	
186	4	5,7692	5,2727	1,1355	[4,6524, 5,8931]	Conectivas	54,55%	Marcado: contenido
187	6	6,7647	7,1333	1,0242	[6,6677, 7,5990]	Sintaxis	100,00%	
188								Retirado: C.it-2
189	8	9,0625	9,3333	1,2472	[8,6867, 9,9800]	Sintaxis	61,54%	
190	4	5,6667	4,6364	0,8814	[4,1548, 5,1179]	Sintaxis	78,57%	
191								Retirado: C.it-2
192	8	8,8667	9	1,3009	[8,3571, 9,6429]	Verbos	75,00%	
193	6	6,4375	5,7692	1,0491	[5,2507, 6,2877]	Sintaxis	70,00%	
194	8	9,3333	8,9231	1,3279	[8,2668, 9,5794]	Verbos	93,33%	
195	5	5,3000	4,8889	0,9938	[4,2727, 5,5050]	Sintaxis	80,00%	
196	8	8,7857	8,75	1,0104	[8,2262, 9,2738]	Verbos	100,00%	
197	7	6,0625	5,8667	1,0873	[5,3723, 6,3610]	Sufijos	53,85%	
198	4	5,3333	4,8571	1,1249	[4,3247, 5,3896]	Declinación	76,92%	Marcado: contenido
199								Retirado: C.it-2
200								Retirado: C.it-2
201								Retirado: C.it-2
202								Retirado: C.it-2
203	9	9,7857	10,4545	1,3048	[9,7417, 11,1674]	Verbos	100,00%	
204								Retirado: C.it-2
205								Retirado: C.it-1
206	4	5,6154	5,3636	1,1499	[4,7354, 5,9919]	Vocabulario	58,33%	
207								Retirado: C.it-2
208								Retirado: C.it-2
209								Retirado: C.it-2
210	5	5,3846	6,3	1,2689	[5,5645, 7,0355]	Sintaxis	61,54%	
211								Retirado: C.it-2
212	8	8,8750	8,8571	1,1249	[8,3247, 9,3896]	Verbos	100,00%	
213	12	10,7143	11,3333	0,9428	[10,8445, 11,8221]	Verbos	100,00%	
214	9	8,7857	8,4545	1,0757	[7,8669, 9,0422]	Verbos	92,86%	
215	7	7,6667	7,6667	1,1926	[7,1244, 8,2089]	Verbos	100,00%	
216	8	8,0588	7,6667	1,0111	[7,2070, 8,1264]	Verbos	93,75%	
217	6	6,4706	5,7143	0,9583	[5,2607, 6,1679]	Sintaxis	78,57%	
218								Retirado: C.it-2
219	8	8,6000	8,3333	1,0274	[7,8007, 8,8660]	Verbos	100,00%	
220	5	5,9231	5,5833	1,0375	[5,0454, 6,1212]	Declinación	90,00%	

	Dificultad					Destreza		
Item	**moda**	**media**	**M.dif**	**σ(M.dif)**	**IC95 (M.dif)**	**M.est**	**%Moda**	**Detalle**
221	4	5,3810	4,6471	1,0632	[4,3964, 4,8978]	Verbos	59,62%	
222	8	7,8667	7,8667	1,2037	[7,3194, 8,4140]	Verbos	100,00%	
223	8	8,2000	8	0,9129	[7,5267, 8,4733]	Verbos	100,00%	
224	8	9,0000	8,6154	1,003	[8,1197, 9,1111]	Verbos	100,00%	
225								Retirado: C.it-2
226	10	8,7333	8,5833	1,1873	[7,9678, 9,1989]	Sintaxis	36,36%	
227	12	11,6667	11,6667	0,5963	[11,3955, 11,9378]	Verbos	84,62%	
228	6	6,2000	6,3571	0,895	[5,9335, 6,7808]	Sintaxis	100,00%	
229	8	7,7333	7,5833	1,1149	[7,0053, 8,1614]	Sufijos	77,78%	
230	6	8,5333	7,5455	1,1571	[6,9133, 8,1776]	Declinación	100,00%	Marcado: contenido
231	8	8,8125	8,6154	1,1461	[8,0489, 9,1818]	Declinación	81,25%	
232								Retirado: C.it-2
233								Retirado: C.it-2
234								Retirado: C.it-1
235								Retirado: C.it-2
236								Retirado: C.it-2
237								Retirado: C.it-2
238	12	10,6667	11,0769	0,997	[10,5841, 11,5697]	Sintaxis	69,23%	
239								Retirado: C.it-1
240								Retirado: C.it-2
241	7	7,0169	6,587	1,0545	[6,3251, 6,8488]	Declinación	62,50%	
242								Retirado: C.it-2
243								Retirado: C.it-2
244	9	9,7333	9,3846	1,003	[8,8889, 9,8803]	Verbos	100,00%	
245	6	6,4375	6,75	1,0104	[6,2262, 7,2738]	Sintaxis	75,00%	
246	5	6,3077	6	1,1282	[5,3836, 6,6164]	Sintaxis	90,00%	
247	6	6,5263	6,7143	1,0302	[6,2267, 7,2019]	Sintaxis	35,29%	
248								Retirado: C.it-2
249								Retirado: C.it-1
250	8	7,7273	7,1111	0,8749	[6,5687, 7,6535]	Sintaxis	90,00%	
251	3	3,5714	2,9167	0,6401	[2,5848, 3,2485]	Vocabulario	80,00%	
252	7	7,8125	6,9231	0,9166	[6,4700, 7,3761]	Declinación	54,55%	

Tabla 2.- Dificultad y destreza estimada de los ítems (m=3315>2876, n=192, e=111)

Tras la calibración quedan 9 ítems marcados por contenido de los 22 identificados durante el filtrado del banco que se había realizado con anterioridad a la recogida de datos. El resto (13) fueron eliminados durante el cribado de la muestra por los filtros C.it.

Las siguientes dos figuras confrontan la Moda vs Media de los 192 pares de valores de dificultad estimados (Figura 1 y Figura 2).

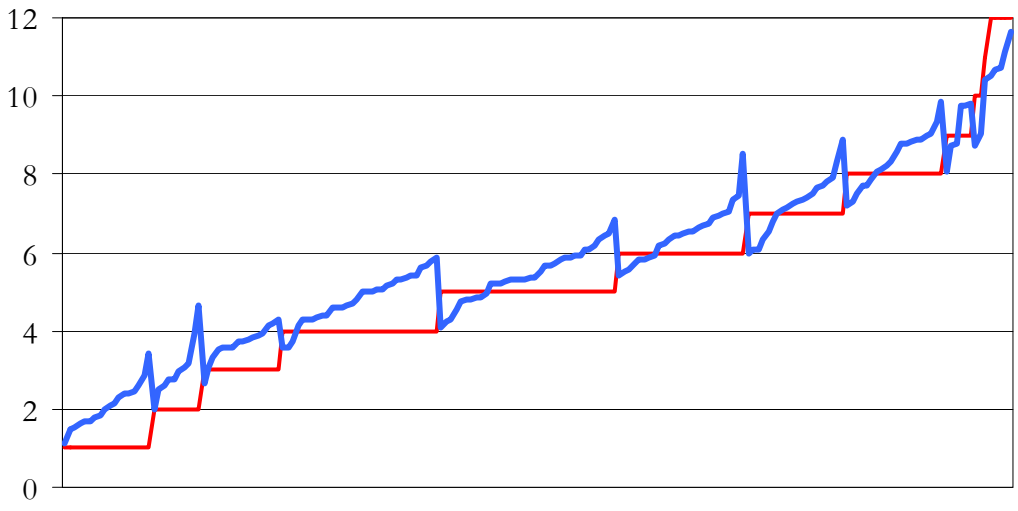


Figura 1.- Contraste de los 192 pares valores de dificultad estimados por los estadísticos Moda y Media, ordenados por valores de crecientes de Moda (en rojo)

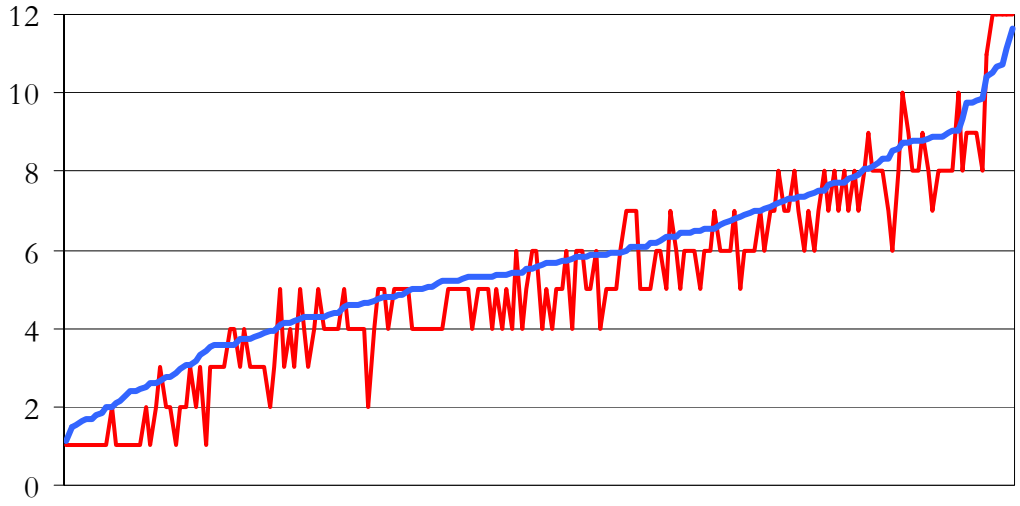


Figura 2.- Contraste de los 192 pares valores de dificultad estimados por los estadísticos Moda y Media, ordenados por valores de crecientes de Media (en azul)

Las siguientes dos figuras confrontan la Moda vs M.dif de los 192 pares de valores de dificultad estimados (Figura 3 y Figura 4).

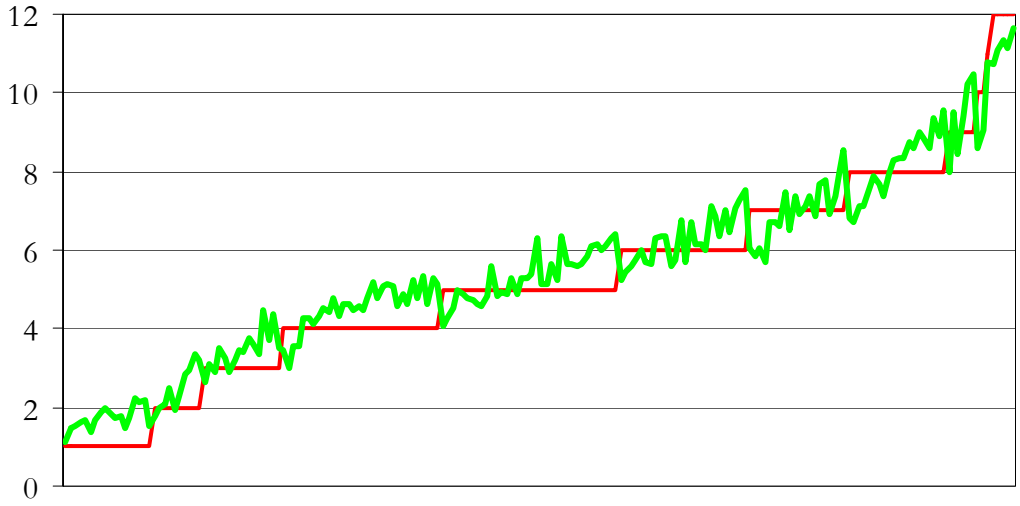


Figura 3.- Contraste de los 192 pares valores de dificultad estimados por los estadísticos Moda y M.dif, ordenados por valores de crecientes de Moda (en rojo)

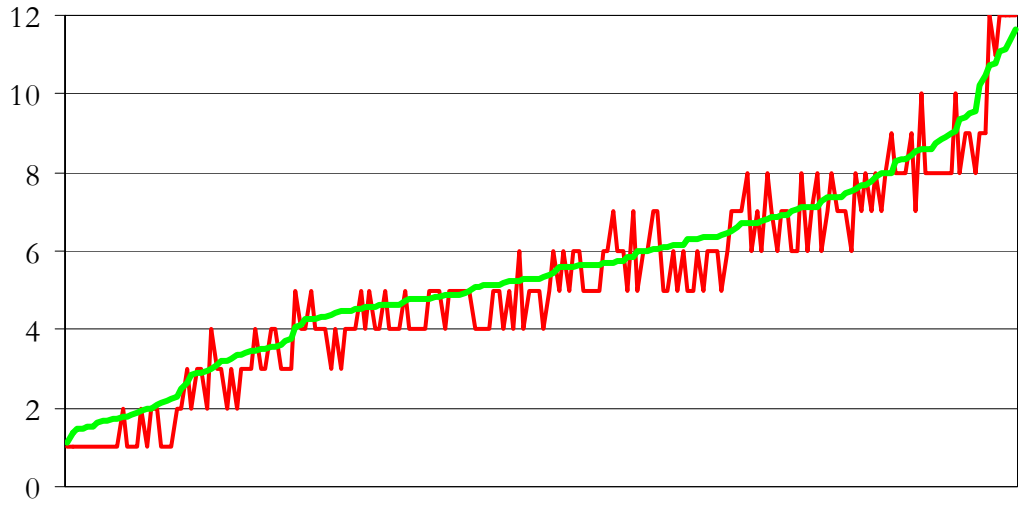


Figura 4.- Contraste de los 192 pares valores de dificultad estimados por los estadísticos Moda y M.dif, ordenados por valores de crecientes de M.dif (en verde)

Las siguientes dos figuras confrontan la Media vs M.dif de los 192 pares de valores de dificultad estimados (Figura 5 y Figura 6).

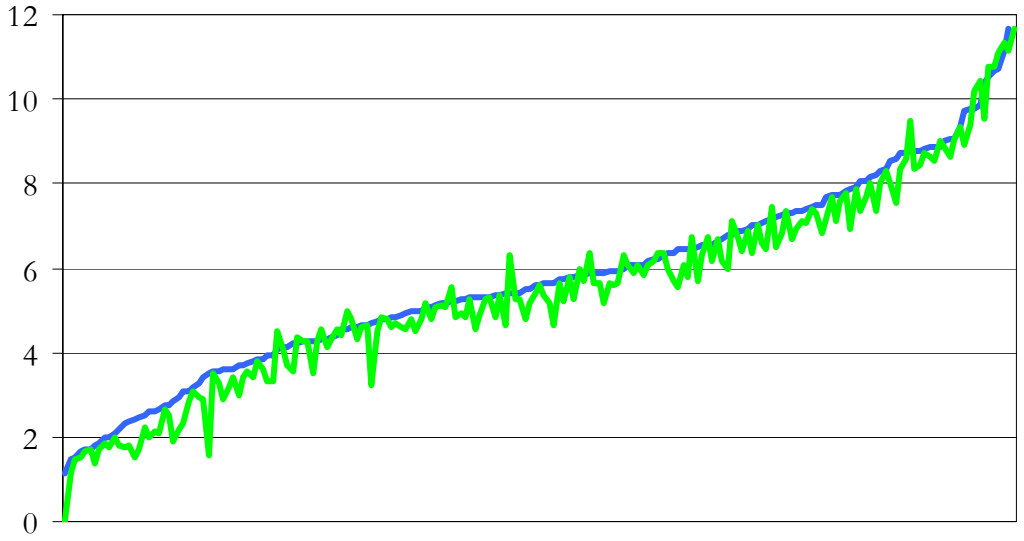


Figura 5.- Contraste de los 192 pares valores de dificultad estimados por los estadísticos Media y M.dif, ordenados por valores de crecientes de Media (en azul)

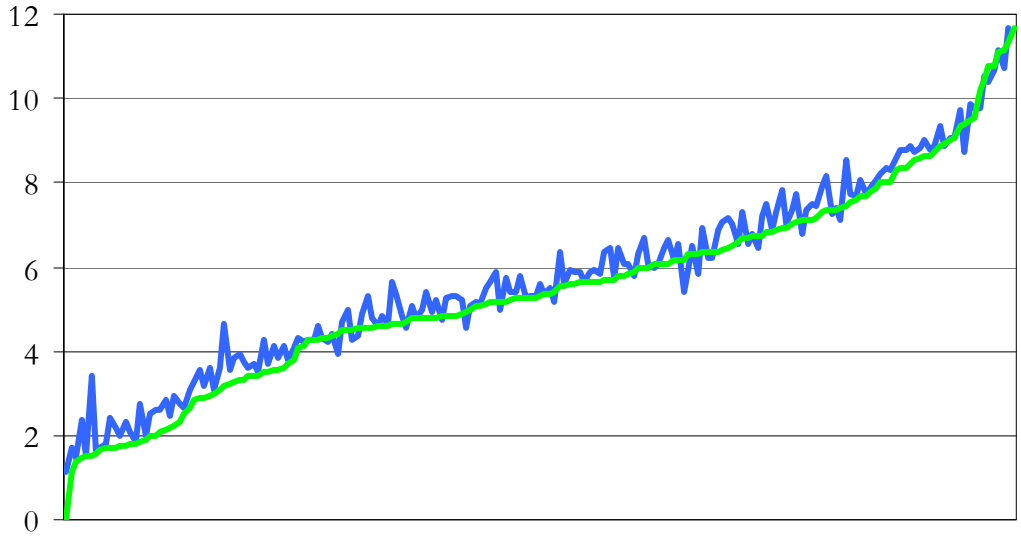


Figura 6.- Contraste de los 192 pares valores de dificultad estimados por los estadísticos Media y M.dif, ordenados por valores de crecientes de M.dif (en verde)

3 Fiabilidad de los resultados

Para dar la validez a los resultados, se minimiza el posible sesgo de evaluadores, se cuantifica el grado de acuerdo usando el test de Kappa. Sin embargo, y aunque coeficiente Kappa es el estimador más empleado para medir la magnitud de la concordancia, hay que recordar que la estimación puntual del valor de **κ** no proporciona ninguna indicación de la precisión de la estimación. En consecuencia, y para conocer la variabilidad de las estimaciones, se ha optado por calcular los cruces entre los intervalos de confianza (IC) para $Mdif_1$ y para $Mdif_2$, al considerar éste un proceso similar pero más intuitivo que plantear un test de hipótesis con intervalos de confianza para Kappa.

Para *minimizar el posible sesgo entre los evaluadores*, en la etapa de diseño de los cuestionarios se optó por incluir en los cuestionarios una tabla de niveles de dificultad equivalentes y reconocidos entre los distintos organismos de la Comunidad Autónoma Vasca, HABE, Gobierno Vasco, IVAP, EOI... Los expertos participantes en el estudio, además de su capacitación lingüística, se hallaban habituados a trabajar con ítems organizados en alguna de las escalas incluidas en la tabla de equivalencias proporcionada.

Para cuantificar el *grado de acuerdo* entre evaluadores, se ha empleado el test de Kappa. En concreto, para el aspecto **Dificultad** del ítem y con escala ordinal [1..12] se ha obtenido un índice de confiabilidad Kappa (**κ**) de **0.675**, lo que significa, como se ha comentado en el capítulo 4, una fuerza de concordancia buena. En el rasgo **Destreza**, el grado de acuerdo entre los dos grupos de expertos es aún mayor, siendo **0.763** el valor concreto del índice de Kappa (**κ**) alcanzado. En la Tabla 3 se han recogido los índices de

confiabilidad ponderados con pesos cuadráticos de cada uno de los ítems en cuanto al rasgo Dificultad (columna *κ-cuád.(Dificultad)*) y el índice Kappa para el rasgo Destreza (columna *κ(Destreza)*).

En el caso de los juicios relativos al parámetro Destreza, indicar que se han podido calcular los índices Kappa de 124 ítems de los 192 que quedaron en el banco tras el filtrado de la muestra. De ellas el 52,5% de las veces (en 65 ocasiones) el acuerdo ha sido total, reflejándos con el valor 1 en columna "*κ(Destreza)*".

En el 93% de los índices no calculados (63 de 68), estando en total acuerdo en una destreza gramatical todos los expertos de una las dos submuestras, en la otra submuestra hubo algún/os expertos que discreparon, y el índice Kappa no pudo calcularse.

Con el rasgo Dificultad también ha sucedido lo mismo.

Ítem	κ-cuád. (Dificultad)	κ (Destreza)
1	0,8436	0,2961
2	0,4006	0,5
3	0,5728	0,7037
4	0,6394	0,7241
5	0,7075	0,261
6		
7	0,8211	
8	0,8016	
9	0,6285	1
11	0,5309	0,3393
12	0,8494	
13	0,8186	0,7034
14	0,5802	0,8
15	0,6735	0,6563
17	0,872	0,2359
18	0,6902	0,4667
19	0,5851	
20		0,4909
21	0,375	0,2936
22	0,8174	0,9383
23	0,3199	0,7037
26	0,4664	
27	0,8638	
28	0,5	1
29	0,7017	
31	0,7695	
32	0,766	
33	0,694	
34	0,2863	1
35	0,5964	1
36	0,4975	
37	0,7993	
38	0,7325	1
39	0,5003	

Ítem	κ-cuád. (Dificultad)	κ (Destreza)
40	0,955	
41	0,9356	
42	0,913	
43	0,6785	1
44	0,8148	0,4968
45	0,4919	
46	0,6266	
47	0,5843	0,072
48	0,6286	1
49		1
52	0,574	
53	0,7143	
54	0,8629	0,5963
55	0,7564	0,3218
57	0,2996	0,541
58	0,5283	0,4444
61	0,6711	
64	0,6275	
65	0,9436	
66	0,6561	
67	0,6433	1
68	0,747	1
69		
70	0,6243	0,9122
71	1	1
72	0,744	1
73	0,7293	1
74	0,63	1
75	0,6771	
76		1
77	0,438	
78	0,8888	1
79	0,9173	
81	0,6418	

Ítem	κ-cuád. (Dificultad)	κ (Destreza)
82	0,6048	1
84	0,8881	1
85	0,7423	1
86	0,8246	1
87	0,5231	1
88	0,6032	1
89	0,7314	
90	0,6667	1
91	0,567	1
92	0,7854	
93	0,8707	
94	0,6537	0,6554
95	0,5708	1
96	0,6847	
97	0,5571	0,8498
98	0,8481	0,4737
99	0,6364	
100	0,6979	0,3295
101	0,6553	0,3607
102	0,6711	1
104	0,9008	
106	0,8134	1
107		
108	0,5484	
109		1
110	0,7206	1
111	0,2821	1
112	0,7153	0,4
113	0,7351	
114	0,7518	1
116	0,6197	
117	0,5727	
118	0,5455	0,5064
119	0,7921	1
120	0,2462	1
121		1
122	0,1333	1
123	0,8253	1
124		
125	0,7243	
126	0,5706	
127	0,8283	0,79
128	0,6503	0,6653
129		
130		1
132	0,9453	1
133	0,6549	1
137	0,8826	1
138	0,5333	1
139	0,8351	1
140	0,6583	
141	0,72	
142	0,4718	0,4655
146		0,5418

Ítem	κ-cuád. (Dificultad)	κ (Destreza)
147	0,8658	
149	0,2962	
150	0,7443	0,3182
152	0,7605	1
153	0,7963	1
154	0,7389	1
155	0,6	1
156	0,6568	1
158	0,2432	1
159	0,4682	1
160	0,7748	1
162	0,8162	0,4279
164	0,8922	1
165	0,7849	1
166	0,5188	1
167		
171	0,505	
172	0,7399	0,5682
173	0,5069	
174		0,3755
175	0,4311	0,0881
177	0,8207	1
179	0,3488	0,1189
180	0,8351	
181	0,8364	0,2438
182	0,6928	1
183	0,5366	0,4605
184		1
185	0,8046	
186	0,9232	0,4164
187	0,6107	1
189	0,876	0
190	0,2222	
192	0,6831	0,4444
193	0,6522	
194	0,8311	
195		
196	0,2708	1
197	0,7266	0,2747
198	0,8606	0,2939
203	0,8665	1
206	0,3985	0,4812
210		0,7996
212	0,695	1
213	0,613	1
214	0,8885	
215		1
216	0,7501	
217	0,3351	0,435
219		1
220	0,8056	
221	0,7356	0,5625
222		1
223	0,7106	1

Ítem	κ-cuád. (Dificultad)	κ (Destreza)
224	0,7479	1
226	0,8327	0,1408
227	0,6032	
228	0,7794	1
229		
230	0,8355	1
231		0,5925
238	0,7729	
241	0,6361	0,4765

Ítem	κ-cuád. (Dificultad)	κ (Destreza)
244	0,7172	1
245	0,8706	0,1898
246		
247	0,6898	0,324
250	0,8372	
251	0,447	
252		
Total	0,675	0,763

Tabla 3.- Desglose de los niveles de acuerdo entre expertos de la PE1 y PE2 en los rasgos Dificultad y Destreza

El cálculo de los coeficientes de Kappa se ha realizado a partir de las distribuciones porcentuales de las respuestas de los expertos de la PE1 frente a los de la PE2; y el software empleado ha sido la calculadora online gratuita **http://faculty.vassar.edu/lowry/kappa.html**.

Para cada ítem *j* se han calculado el nivel de dificultad D_{1j} junto con su respectivo IC con nivel de confianza del 95%, denotado $IC_{1j}95$, y empleando únicamente los pronósticos de los expertos de la PE1 (1880 valoraciones de 2933); e igualmente, se calculan D_{2j} y $IC_{2j}95$ pero con la submuestra de la PE2 (las 1053 valoraciones restantes de m=2933). Indicar que el 98% de los pares de intervalos de confianza se solaparon, siendo $IC_{1j}95 \cap IC_{2j}95 \neq \varnothing$. Esta cifra avala que la estimación de dificultad estimada empleando el estadístico Mdif a partir de los pronósticos conjuntos de los expertos de la PE1 y de la PE2 han dado origen a estimaciones consensuadas de forma off-line, y no han predominado las valoraciones de uno de los dos subgrupos a la hora de determinar un nivel de dificultad. Ordenados por orden creciente de dificultad estimada por los expertos de la PE1, la Figura 7 muestra gráficamente el solapamiento de los IICC, El gráfico corrobora la tendencia similar de las estimaciones de dificultad emitidas en ambas submuestras.

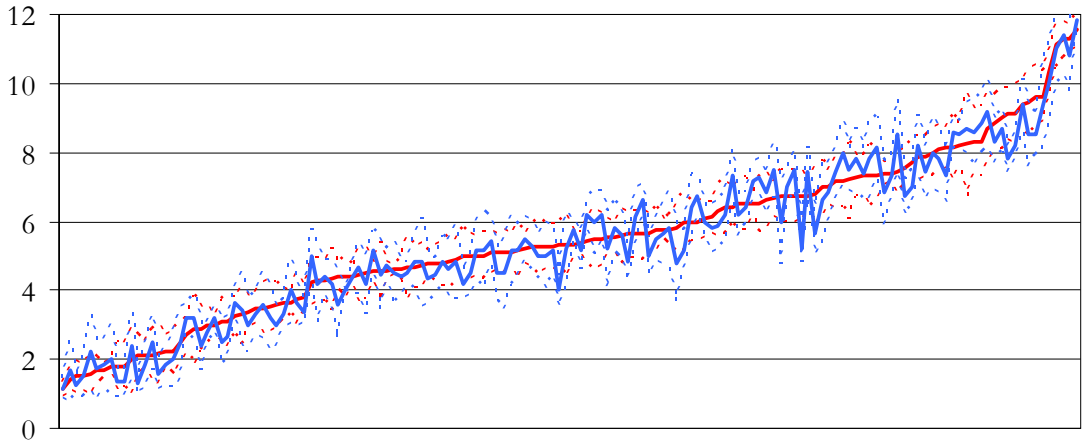


Figura 7.- Solapamiento de los IICC (95%) de los estadístico $M.dif_1$ y $M.dif_2$ entre submuestras PE1 y PE2 (m=1880; 1053, n=192, e=111)

Los valores concretos de los IICC de los estadísticos $M.dif_1$ y para $M.dif_2$, y las ecuaciones de los que se han derivado pueden consultarse en (Arruabarrena & Armendariz, 2008).

Con respecto al parámetro destreza lingüística trabajada por el ítem, se procede de forma análoga y se calculan estimaciones de destrezas por separado, empleando el estadístico M.est con los pronósticos restringidos a los recogidos en submuestra de la PE1 y separadamente con los de la PE2. La Tabla 4 sintetiza los resultados de destreza estimados separadamente y en conjunto. A modo de balance, indicar que difieren en a lo sumo un 1% los volúmenes de las estimaciones de cada tipo de destreza estimada de una variante a otra.

		PE1	**PE2**	**PE1&2**
Destreza	Verbos	34,9%	35,3%	34,9%
	Declinación	18,8%	20,5%	18,2%
	Sintaxis	35,9%	35,3%	37%
	Vocabulario	2,6%	2,1%	2,1%
	Ortografía	0%	0%	0%
	Conectivas	3,1%	3,2%	3,1%
	Sufijos	2,6%	3,7%	3,6%
	Exp. escrita	2,1%	0%	1%

Tabla 4.- Distribución de los volúmenes de destrezas estimadas de los ítems con aportaciones de PE1, de PE2 y de PE1&2 (m= 1859; 1017; 2876, n=192, e=111)

En el documento (Arruabarrena & Armendariz, 2008) se pueden consultar los valores concretos de las estimaciones de destrezas de los ítems empleando las dos submuestras disjuntas.

ANEXO 4 Cuestionario electrónico de la CT

En las pruebas de campo PT1 y PT2 los 250 ítems del banco se distribuyen en 6 cuestionarios. Cada cuestionario contiene 60 ítems, de los cuales los 22 primeros son ítems de anclaje y el resto específicos del cuestionario.

La estructura del cuestionario de las pruebas con expertos se adapta para ser presentado en el ordenador transformándose en una secuencia de sucesivas pantallas. Toda la información se presenta en bilingüe (euskera y castellano). La estructura consta de una *identificación*, la *introducción*, una *recogida de datos personales*, *instrucciones* para completar el test y uno de los 6 *subtests* finalizando con la *puntuación alcanzada*.

La pantalla de *identificación* solicita una *clave de identificación de sesión* (que no de acceso al sistema). Dicha clave es una elección del sujeto, aunque se sugiere que sea algún método de contacto con él para verificar la validez de su contribución, como una dirección de correo electrónico o un teléfono.

La *introducción* presenta el objetivo del trabajo y las instrucciones de rellenado del cuestionario ilustradas con ejemplos.

La *recogida de datos personales* del participante obtiene algunos datos del experto con fines estadísticos como la edad, sexo, titulación, experiencia con el euskera, localización.

Las *instrucciones* indican cómo responder al subtest utilizando 2 ítems de entrenamiento como ejemplo.

La parte central del cuestionario se dedica a administrar un *subconjunto de los ítems a valorar***,** el correspondiente a uno de los subtests. El sujeto tiene que responder mediante una elección entre 4 opciones presentadas, o bien omitir la respuesta.

La prueba finaliza con una pantalla donde se le indican la *puntuación alcanzada***.** Esta se presenta de dos maneras: *absoluta* que indica el porcentaje de ítems acertados de los 60 presentados y ajustada que ofrece una puntuación en la que cada ítems correcto vale 1pto y cada respuesta incorrecta resta 1/3puntos (el número de opciones incorrectas que hubiera).

Seguidamente, en **el primer apartado** se muestra la **secuenciación de** las **pantallas** mencionadas. En el **segundo apartado** se añaden las versiones en castellano de los **correos electrónicos remitidos**. El primero, bilingüe (euskera y castellano), enviado para la captación de los sujetos anónimos en las sesiones no supervisadas (las PT1), y el segundo, para aquellos que se aportaron su email como forma de contacto, la solicitud de la confirmación de validez de su correspondiente sesión.

1 Pantallas representativas del cuestionario electrónico

Figura 1.- Pantalla de presentación de la aplicación de administración de subtest

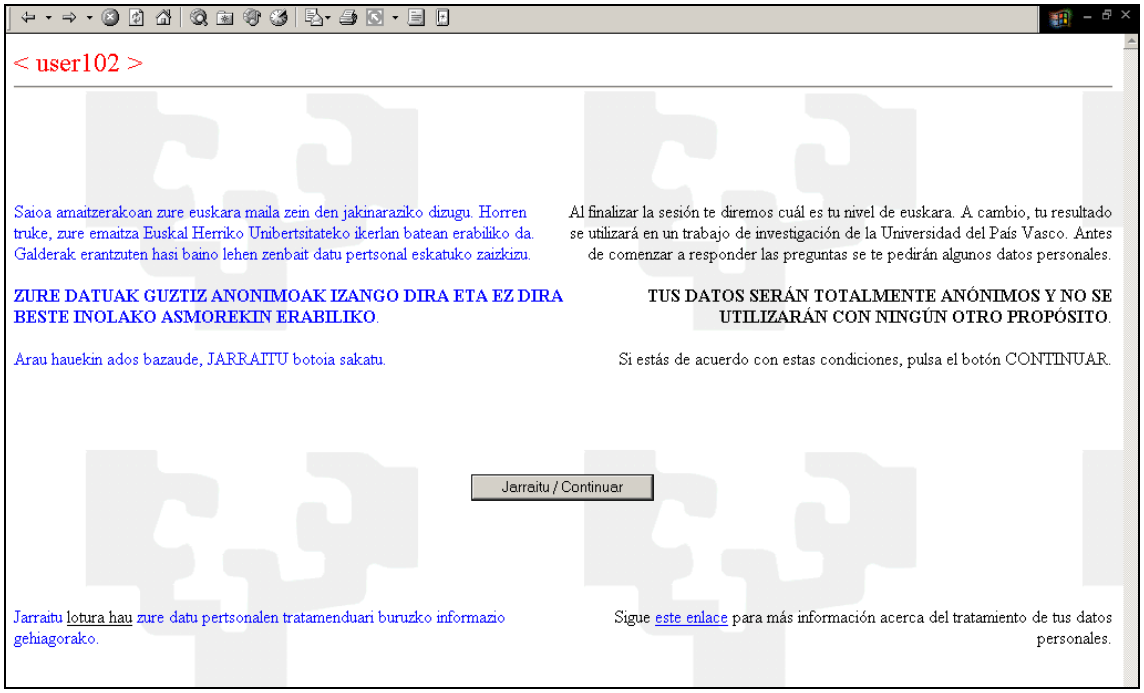
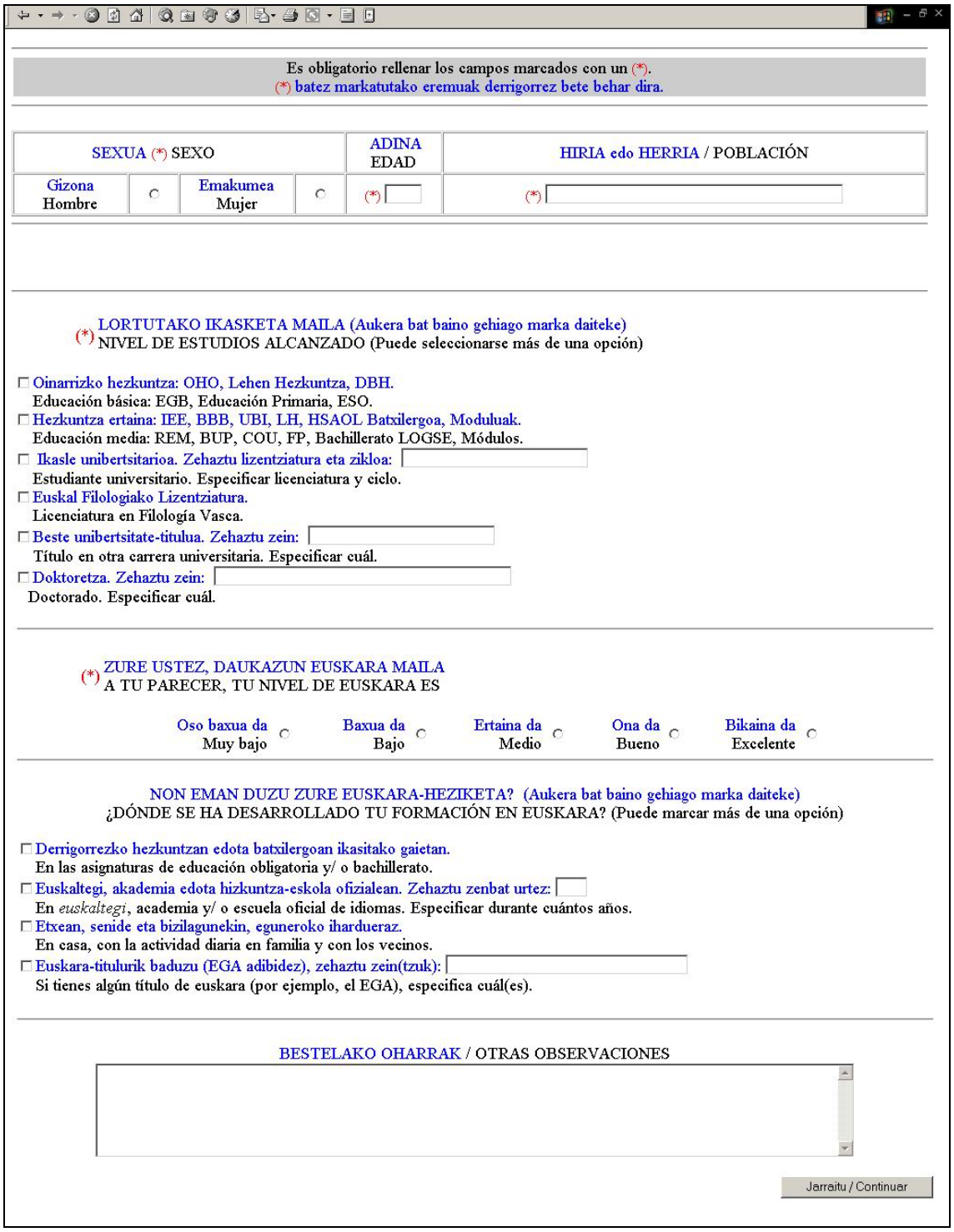


Figura 2.- Primeras instrucciones y aceptación de las condiciones



Es obligatorio rellenar los campos marcados con un (*).
(*) batez markatutako eremuak derrigorrez bete behar dira.

SEXUA (*) SEXO				ADINA EDAD	HIRIA edo HERRIA / POBLACIÓN
Gizona Hombre	○	Emakumea Mujer	○	(*)	(*)

(*) LORTUTAKO IKASKETA MAILA (Aukera bat baino gehiago marka daiteke)
NIVEL DE ESTUDIOS ALCANZADO (Puede seleccionarse más de una opción)

☐ Oinarrizko hezkuntza: OHO, Lehen Hezkuntza, DBH.
Educación básica: EGB, Educación Primaria, ESO.
☐ Hezkuntza ertaina: IEE, BBB, UBI, LH, HSAOL Batxilergoa, Moduluak.
Educación media: REM, BUP, COU, FP, Bachillerato LOGSE, Módulos.
☐ Ikasle unibertsitarioa. Zehaztu lizentziatura eta zikloa:
Estudiante universitario. Especificar licenciatura y ciclo.
☐ Euskal Filologiako Lizentziatura.
Licenciatura en Filología Vasca.
☐ Beste unibertsitate-titulua. Zehaztu zein:
Título en otra carrera universitaria. Especificar cuál.
☐ Doktoretza. Zehaztu zein:
Doctorado. Especificar cuál.

(*) ZURE USTEZ, DAUKAZUN EUSKARA MAILA
A TU PARECER, TU NIVEL DE EUSKARA ES

Oso baxua da Muy bajo ○ Baxua da Bajo ○ Ertaina da Medio ○ Ona da Bueno ○ Bikaina da Excelente ○

NON EMAN DUZU ZURE EUSKARA-HEZIKETA? (Aukera bat baino gehiago marka daiteke)
¿DÓNDE SE HA DESARROLLADO TU FORMACIÓN EN EUSKARA? (Puede marcar más de una opción)

☐ Derrigorrezko hezkuntzan edota batxilergoan ikasitako gaietan.
En las asignaturas de educación obligatoria y/ o bachillerato.
☐ Euskaltegi, akademia edota hizkuntza-eskola ofizialean. Zehaztu zenbat urtez:
En *euskaltegi*, academia y/ o escuela oficial de idiomas. Especificar durante cuántos años.
☐ Etxean, senide eta bizilagunekin, eguneroko iharduerez.
En casa, con la actividad diaria en familia y con los vecinos.
☐ Euskara-titulurik baduzu (EGA adibidez), zehaztu zein(tzuk):
Si tienes algún título de euskara (por ejemplo, el EGA), especifica cuál(es).

BESTELAKO OHARRAK / OTRAS OBSERVACIONES

Jarraitu / Continuar

Figura 3.- Pantalla de petición de datos personales no identificativos

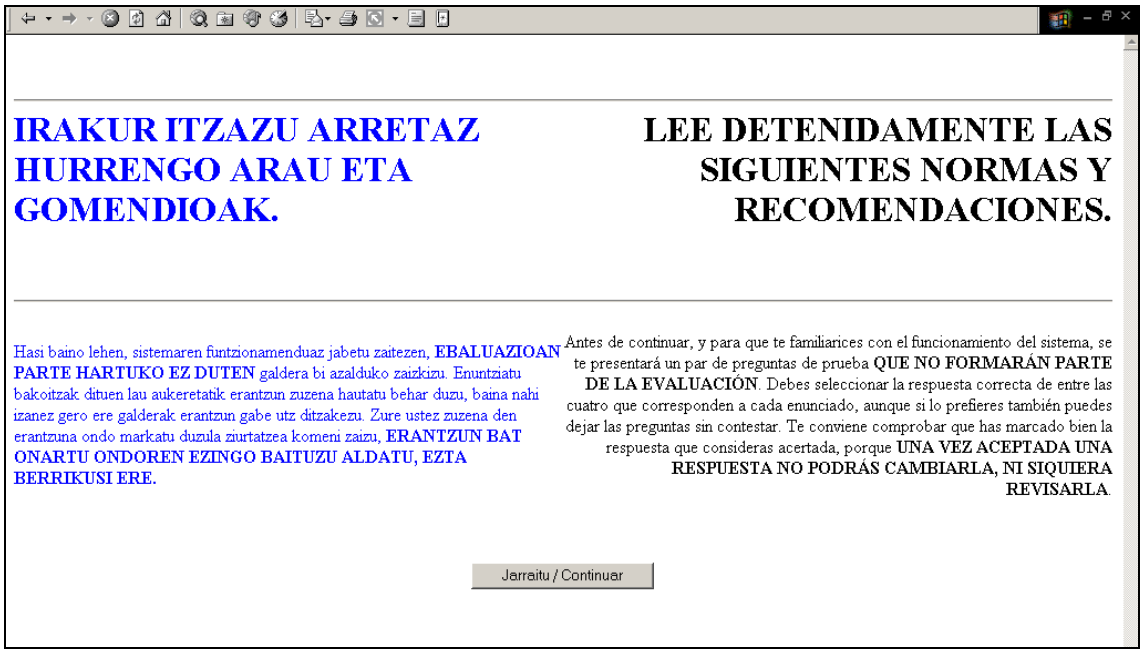


Figura 4.- Primeras instrucciones antes de la administración del subtest

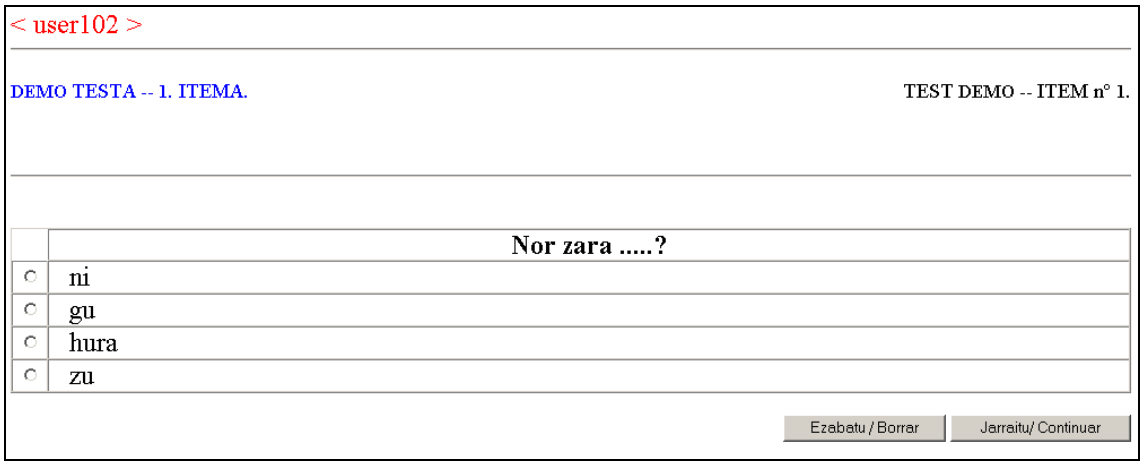


Figura 5.- Pantalla de administración de los ítems de prueba

Figura 6.- Últimas instrucciones antes de la administración del subtest

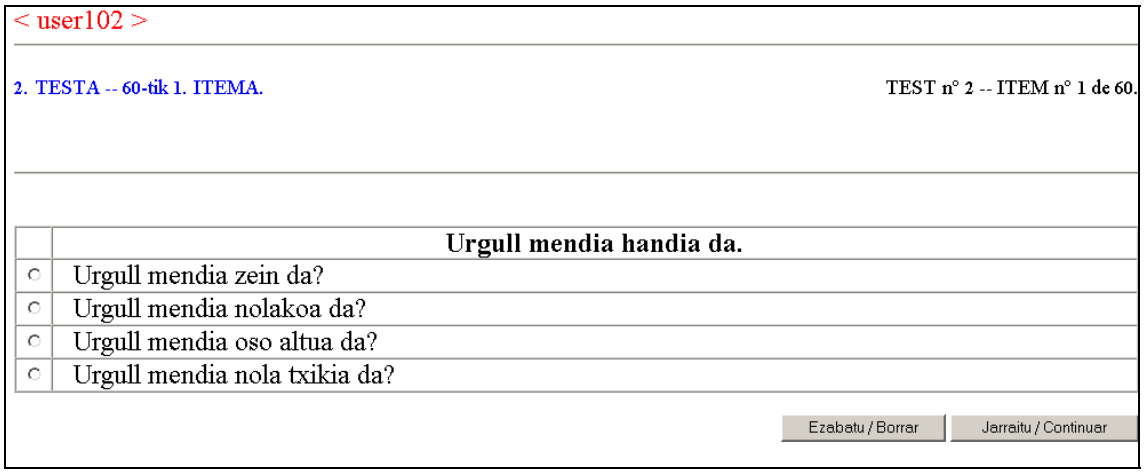


Figura 7.- Administración de los ítems que componen el subtest

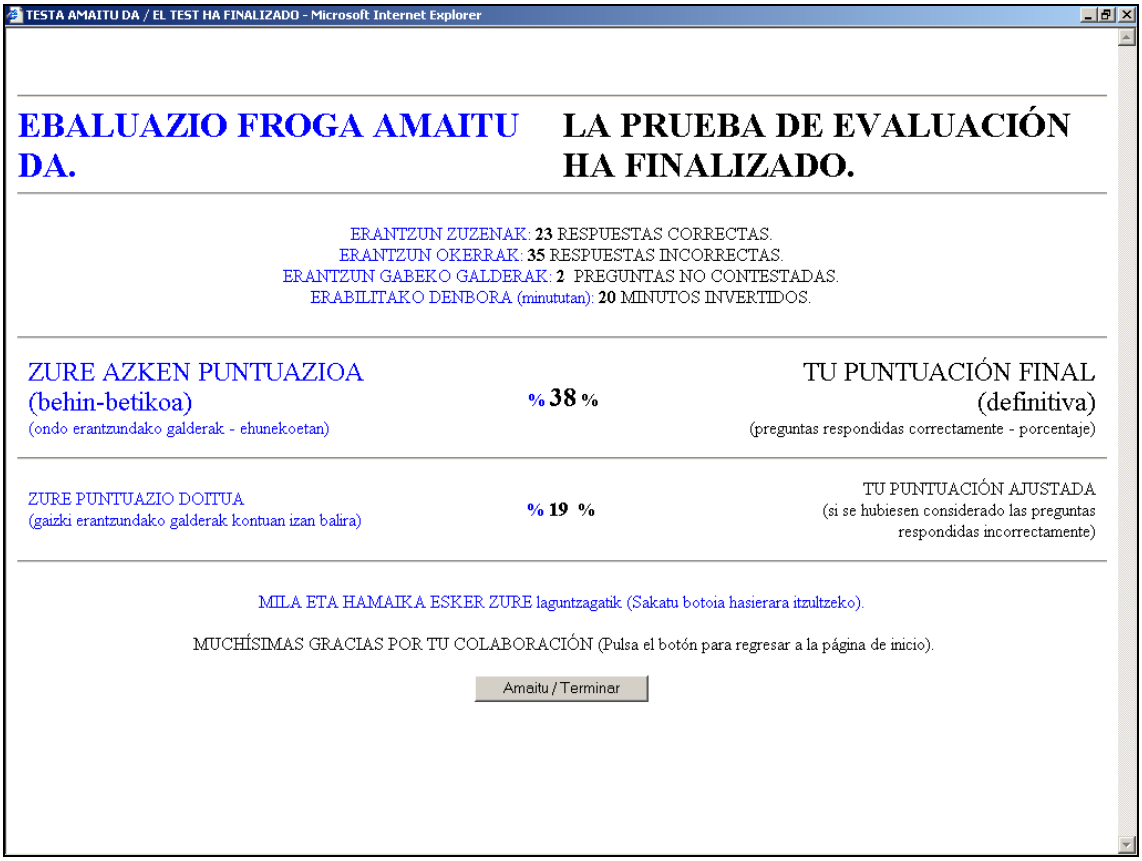


Figura 8.- Pantalla de finalización del subtest

2 Correos electrónicos remitidos en la PT1

Estimado compañero o compañera,

parte del trabajo asociado a nuestras tesis doctorales consiste en administrar una serie de preguntas (ítems) de evaluación del nivel de euskera a una muestra de más de 3000 personas lo más heterogénea posible, con el fin de poder realizar determinados análisis estadísticos a partir de las respuestas. Si estás interesado/a en los detalles, no dudes en solicitarnos más información a la dirección electrónica javilo@si.ehu.es.

Necesitamos personas que sepan nada, poco y mucho euskera, y por este motivo nos dirigimos a ti, para que formes parte de la muestra. Lo único que te pedimos es que te conectes a la dirección web que indicamos al final de este mensaje, y en ella rellenes un test de 60 preguntas. Quienes han realizado las pruebas piloto han tardado menos de media hora en finalizar el test (aunque tienes hasta 50 minutos para hacer la prueba).

En la página de presentación se te pedirá un código de identificación. Basta con que escribas ahí tu dirección de correo electrónico (o número de teléfono). Este detalle es muy importante porque tenemos que ponernos en contacto contigo para que nos confirmes que has realizado el test con seriedad y sin interrupciones (y por tanto aceptar la prueba como válida), o si por el contrario respondiste las preguntas al azar o sin prestar toda tu atención.

Agradeciendo de antemano tu colaboración, y solicitándote asimismo que des a este mensaje la mayor difusión posible (entre los amigos, familiares o conocidos a quienes pueda gustarle la idea de hacer un test de nivel de euskera), recibe un cordial saludo

Javier López-Cuadrado y Rosa Arruabarrena

Grupo de Investigación en Hipermedia y Multimedia, Departamento de Lenguajes y Sistemas Informáticos, Universidad del País Vasco (UPV-EHU)

La dirección web en la que se encuentra el test es http://ji.ehu.es/javilo/

Figura 9.- Versión en castellano del mensaje enviado a diversas listas de distribución

Muchas gracias por tu ayuda,

hemos encontrado esta dirección electrónica en nuestra base de datos, en el campo asociado al "código de identificación".

Es muy importante para nosotros saber si la prueba realizada es válida. Por ello, si el resultado es fiable, te agradeceríamos que nos lo confirmaras. Cuando te preguntamos si la prueba es válida o no, no nos referimos a la puntuación final en el test, sino a las condiciones en las que la desarrollaste: por ejemplo, si no tuviste interrupciones, si no pediste ayuda a otras personas, etc... dado que de otra manera los resultados del test no serían fiables (independientemente de la puntuación final). Recuerda que lo que para ti fue una pequeña prueba de nivel de euskera, para nosotros es información de gran utilidad para el desarrollo de dos tesis doctorales.

BASTA CON QUE RESPONDAS A ESTE MENSAJE (por ejemplo, mediante el comando REPLY) incluyendo la palabra "ONETSI" en el texto de tu respuesta. Si por el contrario, realizaste la prueba por curiosidad, y, por lo tanto, sin afán de obtener una estimación de tu nivel de euskera, perdona las molestias y, por favor, NO RESPONDAS A ESTE MENSAJE.

Queremos agradecer nuevamente tu ayuda; recibe un cordial saludo.

Javier López-Cuadrado y Rosa Arruabarrena

Grupo de Investigación en Hipermedia y Multimedia, Departamento de Lenguajes y Sistemas Informáticos, Universidad del País Vasco (UPV-EHU).

Figura 10.- Mensaje de confirmación de la validez de una sesión no supervisada

ANEXO 5 Resultados de la calibración 3PL-TRI

Una vez filtrada la muestra de los sujetos anónimos de las pruebas PT1 y PT2, la muestra contiene 193.630 valoraciones sobre 204 ítems, valoraciones que se han emitido por 3243 sujetos. Seguidamente, se procede a la estimación de los tres parámetros del modelo logístico de la TRI (3Pl-TRI).

En este anexo se muestran los valores estimados de los parámetros una vez equiparadas las escalas. El primer campo de la tabla siguiente corresponde al identificador del ítem (columna *"Id ítem"*), los tres siguientes recogen los valores equiparados de los parámetros *a*, *b* y *c* de aquellos ítems que superaron los procesos de filtrado, validez y bondad de ajuste (columnas *"Discriminación"*, *"Dificultad"* y *"Pseudoacierto"*) y el último campo indica los motivos de eliminación de los 48 ítems que fueron retirados del banco, o de marcado de los 80 que, pese a permanecer en el banco, fueron señalados como potencialmente erróneos.

Ítem	a_i Discriminación	b_i Dificultad	c_i Pseudoacierto	Detalle
1	---	---	---	Retirado: análisis de fiabilidad
2	---	---	---	Retirado: análisis de fiabilidad
3	---	---	---	Retirado: análisis de fiabilidad
4	---	---	---	Retirado: análisis de fiabilidad
5	---	---	---	Retirado: análisis de fiabilidad
6	---	---	---	Retirado: análisis de fiabilidad
7	1,052885187	-2,848497669	0,23	
8	0,837797544	-2,066336145	0,23	Marcado: funcionamiento diferencial
9	0,685158238	-1,283520504	0,22	Marcado: distractor nunca elegido y escasa correlación
10	0,895440234	-1,934049183	0,22	
11	---	---	---	Retirado: análisis de fiabilidad
12	---	---	---	Retirado: análisis de fiabilidad
13	0,694753556	-0,643550273	0,22	Marcado: contenido y diferencias en porcentaje (SV/ NSV)
14	---	---	---	Retirado: análisis de fiabilidad
15	0,991206399	-1,581964072	0,24	
16	---	---	---	Retirado: análisis de fiabilidad
17	1,036825534	-2,273546994	0,22	
18	---	---	---	Retirado: análisis de fiabilidad
19	0,846174203	-0,407783218	0,27	Marcado: correlación escasa, diferencias en porcentaje y FDI
20	0,703660653	-2,035698599	0,23	
21	---	---	---	Retirado: análisis de fiabilidad
22	1,21692284	-1,332414319	0,22	
23	---	---	---	Retirado: análisis de fiabilidad
24	0,757103234	0,063750893	0,24	Marcado: distractor nunca elegido, diferencias en porcentaje y FDI
25	---	---	---	Retirado: análisis de fiabilidad
26	---	---	---	Retirado: análisis de fiabilidad
27	0,66	-1,54	0,15	Marcado: funcionamiento diferencial
28	0,63	-2,15	0,15	Marcado: correlación escasa y FDI
29	0,841992532	-1,619494096	0,24	
30	1,176017871	-1,977812675	0,22	
31	1,043399735	-3,05934687	0,23	
32	1,063529205	-2,554759689	0,22	
33	0,947509127	-2,788228928	0,22	
34	0,807873146	-0,794582356	0,22	Marcado: distractor nunca elegido
35	1,102805341	-1,541504838	0,21	
36	0,806263432	-2,700903229	0,23	Marcado: correlación escasa
37	1,108445866	-2,191826963	0,23	Marcado: distractor nunca elegido
38	0,927561566	-2,437308825	0,22	
39	1,135112901	-2,466750823	0,22	
40	0,961966462	-1,148765391	0,24	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
41	1,165791628	-2,672104845	0,22	
42	0,92957431	-2,363544508	0,22	
43	1,319185264	-0,589228334	0,22	
44	0,873966932	-1,262958867	0,23	
45	1,306565217	-2,276888207	0,22	
46	0,848778115	-2,662326082	0,22	Marcado: correlación escasa
47	1,253616328	-2,453905207	0,21	
48	1,086665818	-2,372508678	0,22	
49	0,8692306	-2,408078245	0,22	
50	0,820034741	0,834979844	0,21	Marcado: destreza lingüística
51	0,935245172	-1,844839554	0,23	

Ítem	a_i Discriminación	b_i Dificultad	c_i Pseudoacierto	Detalle
52	0,828360009	-2,933858809	0,23	Marcado: correlación escasa
53	1,183052799	-2,829837372	0,23	Marcado: distractor nunca elegido y escasa correlación
54	1,137010952	-2,096415011	0,22	
55	1,065813332	-2,520214673	0,23	
56	---	---	---	Retirado: análisis de fiabilidad
57	---	---	---	Retirado: análisis de fiabilidad
58	0,99759485	-2,945085812	0,23	Marcado: destreza lingüística
59	0,706926501	-2,125016702	0,22	Marcado: contenido
60	1,204369065	-2,088619397	0,23	
61	0,774917428	-2,271465654	0,23	
62	1,053302963	-2,437414534	0,22	
63	0,68819084	-1,003549549	0,22	
64	1,438847999	-1,506904024	0,23	
65	0,967516118	-1,056279465	0,24	
66	0,71811218	-1,264233054	0,24	Marcado: correlación escasa, diferencias en porcentaje y FDI
67	1,187636521	-1,339928012	0,21	
68	1,042130334	-0,564961254	0,23	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
69	0,926338075	-2,035698599	0,22	
70	1,1097979	-1,541232626	0,23	Marcado: destreza lingüística
71	1,34	-2,30	0,15	
72	1,113387109	-1,945882578	0,22	
73	0,692438006	-0,465901703	0,24	Marcado: correlación escasa y diferencias en porcentajes
74	0,914291607	-1,074695347	0,22	
75	0,999122788	-2,220500461	0,22	
76	1,289634132	-0,165205664	0,24	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
77	1,131201303	-2,013244593	0,22	Marcado: contenido
78	1,247601567	-1,615998445	0,21	
79	1,186879854	-1,73546862	0,22	
80	0,986487022	-0,940312405	0,23	
81	0,989697101	-0,798853375	0,23	
82	---	---	---	Retirado: análisis de fiabilidad
83	0,8550813	-1,406986452	0,23	
84	0,93	-1,10	0,13	
85	0,93	-1,81	0,15	Marcado: funcionamiento diferencial
86	1,06719449	-0,943391817	0,23	
87	1,008548474	-2,167453928	0,22	Marcado: distractor nunca elegido
88	1,1097979	-1,520147706	0,22	
89	1,052885187	-2,268662368	0,22	Marcado: funcionamiento diferencial
90	0,979780656	-1,384532446	0,22	
91	0,807873146	-0,481661941	0,22	
92	1,33	-0,86	0,15	Marcado: funcionamiento diferencial
93	1,185681517	-1,235501285	0,23	
94	1,278280294	-1,146617823	0,22	
95	1,027399848	-2,390249367	0,22	
96	1,138254257	-1,224958825	0,23	
97	0,73628945	-1,948476386	0,22	Marcado: distractor nunca elegido
98	0,905941332	-1,638259108	0,23	Marcado: destreza lingüística
99	1,299506943	-1,625572306	0,21	
100	0,93	-0,67	0,14	
101	1,264807756	-1,406986452	0,22	
102	1,290021491	-1,383095726	0,21	Marcado: contenido
103	1,196470356	-2,251618038	0,22	Marcado: distractor nunca elegido

Ítem	a_i Discriminación	b_i Dificultad	c_i Pseudoacierto	Detalle
104	1,342924799	-0,653095976	0,22	
105	0,927561566	-0,361867076	0,21	
106	0,837797544	-1,414627383	0,22	Marcado: funcionamiento diferencial
107	1,065813332	-2,820454866	0,23	Marcado: correlación escasa
108	0,728085961	-0,442077385	0,22	
109	0,948573866	-0,756303543	0,23	
110	1,24	-1,49	0,15	Marcado: funcionamiento diferencial
111	0,837797544	-0,853155219	0,22	
112	0,987404248	-2,978728412	0,22	Marcado: correlación escasa
113	1,299506943	-1,477977866	0,23	
114	1,127037172	-2,387177382	0,22	
115	0,947509127	-2,136520166	0,22	
116	1,204369065	-2,088619397	0,23	
117	0,920088857	-2,205407608	0,23	
118	0,72	-0,22	0,13	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
119	0,72	-0,77	0,14	
120	1,093379654	-1,308100092	0,21	
121	---	---	---	Retirado: análisis de fiabilidad
122	1,04724693	-1,063707281	0,22	
123	---	---	---	Retirado: análisis de fiabilidad
124	1,055155199	-2,060471878	0,24	Marcado: distractor nunca elegido
125	1,244190641	-1,011039507	0,20	Marcado: destreza y diferencias en porcentaje (SV/ NSV)
126	1,27	-0,51	0,14	
127	1,187636521	-1,329318705	0,21	
128	0,852650666	-0,73753853	0,23	
129	---	---	---	Retirado: análisis de fiabilidad
130	1,406873599	-0,756303543	0,23	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
131	1,306565217	-0,081130994	0,19	
132	0,848311801	-0,607885856	0,22	
133	1,06	-1,24	0,15	
134	1,290021491	-1,615029846	0,23	
135	---	---	---	Retirado: análisis de fiabilidad
136	1,017325589	-1,444706249	0,22	
137	1,280536039	-0,750548124	0,21	
138	0,997378028	-0,33178821	0,21	Marcado: diferencias en porcentajes (SV/ NSV)
139	0,8550813	-0,250605181	0,23	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
140	1,289634132	-0,868893615	0,23	
141	1,202458078	-2,664410746	0,22	
142	1,206696598	-1,469317	0,21	
143	0,834719788	-1,825879047	0,22	
144	0,810609054	-1,223225639	0,22	
145	---	---	---	Retirado: análisis de fiabilidad
146	0,831334399	-2,370094577	0,24	Marcado: contenido y correlacionar escasamente
147	0,819452912	-1,474348467	0,24	Marcado: funcionamiento diferencial
148	---	---	---	Retirado: análisis de fiabilidad
149	0,889683085	-1,66489226	0,22	
150	1,05722071	-0,391945942	0,20	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
151	---	---	---	Retirado: análisis de fiabilidad
152	1,24259423	-0,813802884	0,21	Marcado: contenido
153	---	---	---	Retirado: análisis de fiabilidad
154	0,825234336	-0,908685025	0,23	

Ítem	a_i Discriminación	b_i Dificultad	c_i Pseudoacierto	Detalle
155	1,114660417	1,532763229	0,17	Marcado: correlación escasa, diferencias en porcentaje y FDI
156	0,766968177	0,486435592	0,21	
157	1,071856092	0,504004619	0,22	
158	0,935245172	-1,586618494	0,23	
159	---	---	---	Retirado: análisis de fiabilidad
160	1,380542718	0,682010852	0,16	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
161	---	---	---	Retirado: análisis de fiabilidad
162	1,271050587	-0,560783844	0,20	
163	1,04307672	-0,863033697	0,21	
164	1,087142051	-1,224127899	0,22	
165	0,766968177	-2,232060512	0,22	Marcado: correlación escasa
166	0,937915732	-0,108910628	0,21	
167	1,055155199	-1,112838771	0,23	
168	---	---	---	Retirado: análisis de fiabilidad
169	0,84420524	-0,381562023	0,23	Marcado: destreza lingüística
170	0,778043733	-0,972101181	0,24	Marcado: contenido
171	0,74	-0,76	0,15	
172	1,416228403	-0,396556215	0,21	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
173	---	---	---	Retirado: análisis de fiabilidad
174	0,87	-0,88	0,14	
175	1,097787732	-1,816526722	0,23	Marcado: funcionamiento diferencial
176	1,119103999	-0,540505904	0,23	
177	1,215913581	-0,979211587	0,21	
178	1,175736787	-1,822385549	0,23	Marcado: contenido
179	1,43	0,41	0,11	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
180	1,327157434	-0,014838126	0,20	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
181	1,1097979	-2,057813167	0,22	Marcado: destreza lingüística y FDI
182	---	---	---	Retirado: análisis de fiabilidad
183	0,848778115	-1,224847926	0,22	
184	1,176017871	-1,645334734	0,22	
185	0,887666445	-1,725442331	0,22	
186	1,185681517	-0,655665984	0,20	Marcado: contenido y diferencias en porcentaje (SV/ NSV)
187	0,706926501	0,803151924	0,23	Marcado: correlación escasa
188	0,890709687	1,680439271	0,19	Marcado: destreza, diferencias en porcentajes, escasa correlación y FDI
189	---	---	---	Retirado: análisis de fiabilidad
190	0,8550813	-1,182446399	0,22	
191	---	---	---	Retirado: análisis de fiabilidad
192	1,137010952	-1,304338208	0,22	
193	1,364241065	-1,122221277	0,23	
194	1,095572915	-0,84563632	0,23	
195	0,669223754	-1,732472357	0,22	Marcado: correlación escasa
196	1,505299371	-0,868090325	0,20	
197	1,622012316	-1,024652085	0,20	
198	1,227149083	-0,364316786	0,20	Marcado: contenido y destreza lingüística
199	---	---	---	Retirado: análisis de fiabilidad
200	---	---	---	Retirado: análisis de fiabilidad
201	1,005457927	-1,214416365	0,23	
202	0,801183367	0,230249367	0,24	Marcado: contenido y correlacionar escasamente
203	1,196470356	-0,110068949	0,19	

Ítem	a_i Discriminación	b_i Dificultad	c_i Pseudoacierto	Detalle
204	1,042130334	0,445468982	0,19	Marcado: funcionamiento diferencial
205	---	---	---	Retirado: análisis de fiabilidad
206	0,67	-1,44	0,14	
207	1,300744761	1,63067784	0,14	Marcado: correlación escasa, diferencias en porcentaje y FDI
208	0,930588054	1,542541992	0,19	Marcado: correlación escasa, diferencias en porcentaje y FDI
209	0,961966462	-0,856863323	0,22	
210	1,157225161	-0,392104483	0,22	Marcado: diferencias en porcentajes (SV/ NSV)
211	1,168785148	-0,576057936	0,22	
212	1,310170448	-0,268388044	0,20	
213	---	---	---	Retirado: análisis de fiabilidad
214	1,175736787	-0,463918231	0,20	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
215	1,012522666	-1,150368795	0,23	
216	0,948545214	-0,824345344	0,22	
217	0,644253268	-0,354538023	0,22	
218	0,877692665	-0,141288726	0,21	Marcado: dificultad, diferencias en porcentajes (SV/ NSV) y FDI
219	1,097787732	0,454039733	0,22	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
220	1,236748755	0,179552511	0,18	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
221	0,959231999	0,172564553	0,26	Marcado: correlación escasa, diferencias en porcentaje y FDI
222	1,046251221	0,37877966	0,19	
223	0,99759485	-0,396556215	0,21	
224	1,151078399	0,735514914	0,23	Marcado: correlación escasa
225	---	---	---	Retirado: análisis de fiabilidad
226	---	---	---	Retirado: análisis de fiabilidad
227	---	---	---	Retirado: análisis de fiabilidad
228	1,117063392	-0,792997488	0,22	
229	0,810609054	0,516700646	0,21	Marcado: contenido
230	0,71140891	0,250985578	0,23	Marcado: correlación escasa
231	1,342924799	-0,718773518	0,21	
232	1,128768804	-1,056279465	0,22	
233	0,971493024	0,095285074	0,19	
234	---	---	---	Retirado: análisis de fiabilidad
235	0,841992532	-0,934571157	0,23	
236	---	---	---	Retirado: análisis de fiabilidad
237	0,668032265	-1,395759449	0,23	Marcado: contenido
238	---	---	---	Retirado: análisis de fiabilidad
239	---	---	---	Retirado: análisis de fiabilidad
240	---	---	---	Retirado tras el análisis de expertos
241	1,406873599	-0,371620796	0,22	Marcado: diferencias en porcentaje (SV/ NSV) y FDI
242	1,05722071	0,570577768	0,20	Marcado: contenido, destreza, diferencias en porcentajes y FDI
243	1,065813332	0,529099781	0,22	Marcado: dificultad, diferencias en porcentajes (SV/ NSV) y FDI
244	1,100312448	1,07329746	0,18	
245	0,857737487	-0,109248445	0,21	
246	0,664705753	0,065948785	0,22	Marcado: correlación escasa
247	0,967516118	-0,592411224	0,24	
248	1,508109868	-0,151685671	0,19	
249	---	---	---	Retirado: análisis de fiabilidad
250	---	---	---	Retirado: análisis de fiabilidad

Ítem	a_i Discriminación	b_i Dificultad	c_i Pseudoacierto	Detalle
251	0,801183367	0,304514513	0,24	Marcado: correlación escasa, diferencias en porcentaje y FDI
252	---	---	---	Retirado tras el análisis de expertos

Tabla 1.- Valores de los parámetros a, b y c de los ítems Cal.TRI y causa del retiro o marcaje

ANEXO 6 Valores de contraste: CE vs CT

En este anexo se presentan los contrastes realizados para comparar la CE y la CT. Primeramente se debe concretar una única métrica para las estimaciones de Dificultad de los ítems comunes de los estudios. A continuación se compararán los valores pareados para cada ítem en ambos estudios.

Además de comparar las estimaciones, se pretende también extrapolar los costes temporales y económicos de cada uno de los procesos de calibración.

En el primer apartado se muestran los valores de las estimaciones de dificultad de los 163 ítems comunes a CE y a CT junto con sus correspondientes normalizaciones a una métrica común con valores entre 0 y 1.

En el segundo apartado se recogen los resultados de las pruebas estadística aplicadas a los valores muestrales pareados y normalizados. Concretamente, se han utilizado la prueba de suma de rangos con signos de Wilcoxon y el T-test.

En el tercer apartado se recogen las extrapolaciones de los costes temporales y económicos de la calibración CE para tamaños de bancos alternativos empleando solo aportaciones recabadas a través de pruebas PT2.

En el cuarto apartado se recogen las extrapolaciones de los costes temporales y económicos de la calibración CT para tamaños de bancos alternativos empleando solo aportaciones recabadas a través pruebas PT2.

1 Métrica común para las estimaciones pareadas

La siguiente tabla recoge las estimaciones de los 163 ítems que comunes que han calibrado los procesos CE y CT. Los identificadores originales se hallan en la primera columna (*i*) y una nueva identificación correlativa en la segunda (*k*). Las estimaciones originales computadas por los procesos CE y CT están tabuladas en las columnas D_{ik} y B_{ik}, respectivamente. Los D_{ik} están en el intervalo [1, 12] y los B_{ik} en (-3,5, 3,5). Antes de proceder con las unificación de las escalas al intervalo del 0 al 1, se convierten todos los valores B_{ik} al rango positivo (0,5, 7,5) mediante el cambio de origen $4+ B_{ik}$ comentado en el capítulo 4. Los cuatro procedimientos más habituales de transformación de escalas, comentados en el mismo capítulo se aplican paralelamente a los valores de las columnas D_{ik} y $4+ B_{ik}$, y los nuevos valores obtenidos por las transformaciones por la aplicación del procedimiento de conversión *j* de se recogen dos a dos en las columnas D_k^j y B_k^j.

i	k	D_{ik}	B_{ik}	$4+ B_{ik}$	D_k^1	B_k^1	D_k^2	B_k^2	D_k^3	B_k^3	D_k^4	B_k^4
7	1	1,68852	-2,84850	1,15150	0,15166	0,20812	0,03213	0,04592	0,00186	0,00255	0,02226	0,03079
8	2	2,12500	-2,06634	1,93366	0,19087	0,34949	0,07686	0,21624	0,00234	0,00427	0,02801	0,05170
9	3	2,07143	-1,28352	2,71648	0,18605	0,49098	0,07136	0,38671	0,00228	0,00600	0,02730	0,07263
13	4	2,30769	-0,64355	3,35645	0,20728	0,60665	0,09558	0,52608	0,00254	0,00742	0,03042	0,08974
15	5	1,37500	-1,58196	2,41804	0,12350	0,43704	0,00000	0,32172	0,00151	0,00534	0,01812	0,06465
17	6	4,35714	-2,27355	1,72645	0,39136	0,31204	0,30560	0,17112	0,00479	0,00382	0,05743	0,04616
19	7	1,76923	-0,40778	3,59222	0,15891	0,64926	0,04040	0,57742	0,00194	0,00794	0,02332	0,09605
20	8	4,80000	-2,03570	1,96430	0,43114	0,35503	0,35098	0,22291	0,00528	0,00434	0,06327	0,05252
22	9	2,50000	-1,33241	2,66759	0,22455	0,48214	0,11529	0,37607	0,00275	0,00590	0,03295	0,07132
27	10	1,81481	-1,54000	2,46000	0,16301	0,44462	0,04507	0,33086	0,00199	0,00544	0,02392	0,06577
28	11	1,50000	-2,15000	1,85000	0,13473	0,33437	0,01281	0,19802	0,00165	0,00409	0,01977	0,04946
29	12	2,00000	-1,61949	2,38051	0,17964	0,43026	0,06405	0,31355	0,00220	0,00526	0,02636	0,06365
31	13	2,64706	-3,05935	0,94065	0,23776	0,17002	0,13036	0,00000	0,00291	0,00208	0,03489	0,02515
32	14	3,53846	-2,55476	1,44524	0,31783	0,26121	0,22171	0,10988	0,00389	0,00319	0,04664	0,03864
33	15	2,86667	-2,78823	1,21177	0,25749	0,21902	0,15286	0,05904	0,00315	0,00268	0,03779	0,03240
34	16	4,50000	-0,79458	3,20542	0,40419	0,57935	0,32024	0,49319	0,00495	0,00708	0,05931	0,08570
35	17	4,76923	-1,54150	2,45850	0,42837	0,44435	0,34783	0,33053	0,00524	0,00543	0,06286	0,06573
36	18	2,87500	-2,70090	1,29910	0,25823	0,23480	0,15372	0,07806	0,00316	0,00287	0,03790	0,03473
37	19	4,25000	-2,19183	1,80817	0,38174	0,32681	0,29462	0,18892	0,00467	0,00400	0,05602	0,04835
38	20	3,42857	-2,43731	1,56269	0,30796	0,28244	0,21045	0,13546	0,00377	0,00345	0,04519	0,04178
39	21	1,53846	-2,46675	1,53325	0,13819	0,27712	0,01676	0,12905	0,00169	0,00339	0,02028	0,04100
40	22	4,50000	-1,14877	2,85123	0,40419	0,51534	0,32024	0,41606	0,00495	0,00630	0,05931	0,07623
41	23	3,33333	-2,67210	1,32790	0,29940	0,24001	0,20068	0,08433	0,00366	0,00293	0,04394	0,03550
42	24	3,53846	-2,36354	1,63646	0,31783	0,29578	0,22171	0,15152	0,00389	0,00362	0,04664	0,04375
43	25	5,00000	-0,58923	3,41077	0,44910	0,61647	0,37148	0,53790	0,00550	0,00754	0,06590	0,09120

i	k	$\mathbf{D_{ik}}$	$\mathbf{B_{ik}}$	$4+\mathbf{B_{ik}}$	D_k^1	B_k^1	D_k^2	B_k^2	D_k^3	B_k^3	D_k^4	B_k^4
44	26	3,22222	-1,26296	2,73704	0,28942	0,49470	0,18930	0,39119	0,00354	0,00605	0,04247	0,07318
45	27	3,58824	-2,27689	1,72311	0,32229	0,31144	0,22680	0,17039	0,00394	0,00381	0,04730	0,04607
46	28	3,20000	-2,66233	1,33767	0,28743	0,24177	0,18702	0,08646	0,00352	0,00296	0,04218	0,03577
47	29	3,26786	-2,45391	1,54609	0,29352	0,27944	0,19398	0,13184	0,00359	0,00342	0,04307	0,04134
48	30	3,50000	-2,37251	1,62749	0,31437	0,29416	0,21776	0,14957	0,00385	0,00360	0,04613	0,04351
49	31	4,07692	-2,40808	1,59192	0,36619	0,28773	0,27688	0,14182	0,00448	0,00352	0,05374	0,04256
52	32	3,07692	-2,93386	1,06614	0,27637	0,19270	0,17441	0,02733	0,00338	0,00236	0,04056	0,02851
53	33	3,43750	-2,82984	1,17016	0,30876	0,21150	0,21136	0,04998	0,00378	0,00259	0,04531	0,03129
54	34	3,43750	-2,09642	1,90358	0,30876	0,34406	0,21136	0,20969	0,00378	0,00421	0,04531	0,05090
55	35	2,93333	-2,52021	1,47979	0,26347	0,26746	0,15969	0,11740	0,00322	0,00327	0,03866	0,03957
58	36	4,13333	-2,94509	1,05491	0,37126	0,19067	0,28266	0,02488	0,00454	0,00233	0,05448	0,02821
61	37	4,71429	-2,27147	1,72853	0,42344	0,31242	0,34220	0,17157	0,00518	0,00382	0,06214	0,04622
64	38	3,33333	-1,50690	2,49310	0,29940	0,45061	0,20068	0,33807	0,00366	0,00551	0,04394	0,06666
65	39	3,68750	-1,05628	2,94372	0,33121	0,53205	0,23698	0,43620	0,00405	0,00651	0,04860	0,07871
66	40	3,00000	-1,26423	2,73577	0,26946	0,49447	0,16652	0,39091	0,00330	0,00605	0,03954	0,07315
67	41	4,57143	-1,33993	2,66007	0,41061	0,48079	0,32756	0,37443	0,00502	0,00588	0,06026	0,07112
68	42	4,76923	-0,56496	3,43504	0,42837	0,62085	0,34783	0,54319	0,00524	0,00759	0,06286	0,09184
69	43	4,92308	-2,03570	1,96430	0,44220	0,35503	0,36360	0,22291	0,00541	0,00434	0,06489	0,05252
70	44	4,60000	-1,54123	2,45877	0,41317	0,44440	0,33049	0,33059	0,00506	0,00543	0,06063	0,06574
71	45	4,40000	-2,30000	1,70000	0,39521	0,30726	0,30999	0,16536	0,00484	0,00376	0,05800	0,04545
72	46	4,60606	-1,94588	2,05412	0,41372	0,37126	0,33111	0,24247	0,00506	0,00454	0,06071	0,05492
73	47	4,62500	-0,46590	3,53410	0,41542	0,63876	0,33305	0,56476	0,00508	0,00781	0,06096	0,09449
74	48	4,50000	-1,07470	2,92530	0,40419	0,52872	0,32024	0,43219	0,00495	0,00647	0,05931	0,07821
75	49	4,85714	-2,22050	1,77950	0,43627	0,32163	0,35683	0,18267	0,00534	0,00393	0,06402	0,04758
76	50	4,26667	-0,16521	3,83479	0,38324	0,69311	0,29633	0,63024	0,00469	0,00847	0,05624	0,10253
77	51	3,78571	-2,01324	1,98676	0,34003	0,35909	0,24704	0,22780	0,00416	0,00439	0,04990	0,05312
78	52	4,81250	-1,61600	2,38400	0,43226	0,43089	0,35226	0,31431	0,00529	0,00527	0,06343	0,06374
79	53	4,25000	-1,73547	2,26453	0,38174	0,40929	0,29462	0,28829	0,00467	0,00500	0,05602	0,06055
81	54	4,85714	-0,79885	3,20115	0,43627	0,57858	0,35683	0,49226	0,00534	0,00707	0,06402	0,08559
84	55	5,25000	-1,10000	2,90000	0,47156	0,52415	0,39710	0,42668	0,00577	0,00641	0,06920	0,07754
85	56	5,14286	-1,81000	2,19000	0,46194	0,39582	0,38612	0,27206	0,00565	0,00484	0,06779	0,05855
86	57	5,15385	-0,94339	3,05661	0,46292	0,55246	0,38724	0,46078	0,00566	0,00676	0,06793	0,08173
87	58	4,80000	-2,16745	1,83255	0,43114	0,33122	0,35098	0,19422	0,00528	0,00405	0,06327	0,04900
88	59	4,33333	-1,52015	2,47985	0,38922	0,44821	0,30316	0,33518	0,00476	0,00548	0,05712	0,06630
89	60	5,15385	-2,26866	1,73134	0,46292	0,31292	0,38724	0,17218	0,00566	0,00383	0,06793	0,04629
90	61	4,57143	-1,38453	2,61547	0,41061	0,47272	0,32756	0,36472	0,00502	0,00578	0,06026	0,06993
91	62	5,66667	-0,48166	3,51834	0,50899	0,63591	0,43980	0,56133	0,00623	0,00778	0,07469	0,09407
92	63	4,60714	-0,86000	3,14000	0,41381	0,56753	0,33122	0,47894	0,00506	0,00694	0,06073	0,08396
93	64	5,07143	-1,23550	2,76450	0,45552	0,49966	0,37880	0,39717	0,00557	0,00611	0,06685	0,07392
94	65	5,37500	-1,14662	2,85338	0,48279	0,51572	0,40991	0,41653	0,00591	0,00631	0,07085	0,07629
95	66	4,80000	-2,39025	1,60975	0,43114	0,29095	0,35098	0,14571	0,00528	0,00356	0,06327	0,04304
96	67	7,46667	-1,22496	2,77504	0,67066	0,50157	0,62426	0,39947	0,00821	0,00613	0,09842	0,07420
97	68	5,28571	-1,94848	2,05152	0,47476	0,37080	0,40076	0,24191	0,00581	0,00453	0,06967	0,05485
98	69	5,57143	-1,63826	2,36174	0,50043	0,42686	0,43003	0,30946	0,00612	0,00522	0,07344	0,06315
99	70	5,83333	-1,62557	2,37443	0,52395	0,42916	0,45687	0,31223	0,00641	0,00525	0,07689	0,06349
100	71	5,25000	-0,67000	3,33000	0,47156	0,60187	0,39710	0,52032	0,00577	0,00736	0,06920	0,08904
101	72	5,15385	-1,40699	2,59301	0,46292	0,46867	0,38724	0,35983	0,00566	0,00573	0,06793	0,06933
102	73	6,91667	-1,38310	2,61690	0,62126	0,47298	0,56790	0,36503	0,00760	0,00578	0,09117	0,06997
104	74	6,30769	-0,65310	3,34690	0,56656	0,60492	0,50549	0,52400	0,00693	0,00740	0,08314	0,08949
106	75	5,07692	-1,41463	2,58537	0,45601	0,46728	0,37936	0,35816	0,00558	0,00571	0,06692	0,06913
107	76	5,66667	-2,82045	1,17955	0,50899	0,21319	0,43980	0,05202	0,00623	0,00261	0,07469	0,03154

i	k	D_{ik}	B_{ik}	$4+B_{ik}$	D_{k}^{1}	B_{k}^{1}	D_{k}^{2}	B_{k}^{2}	D_{k}^{3}	B_{k}^{3}	D_{k}^{4}	B_{k}^{4}
108	77	6,06250	-0,44208	3,55792	0,54454	0,64306	0,48036	0,56995	0,00666	0,00786	0,07991	0,09513
109	78	6,07692	-0,75630	3,24370	0,54583	0,58627	0,48184	0,50152	0,00668	0,00717	0,08010	0,08673
110	79	5,28571	-1,49000	2,51000	0,47476	0,45366	0,40076	0,34175	0,00581	0,00555	0,06967	0,06711
111	80	5,57143	-0,85316	3,14684	0,50043	0,56877	0,43003	0,48043	0,00612	0,00695	0,07344	0,08414
112	81	5,64151	-2,97873	1,02127	0,50672	0,18459	0,43722	0,01756	0,00620	0,00226	0,07436	0,02731
113	82	6,00000	-1,47798	2,52202	0,53892	0,45583	0,47396	0,34437	0,00659	0,00557	0,07909	0,06743
114	83	5,42857	-2,38718	1,61282	0,48760	0,29150	0,41540	0,14637	0,00597	0,00356	0,07155	0,04312
116	84	6,35294	-2,08862	1,91138	0,57062	0,34547	0,51012	0,21139	0,00698	0,00422	0,08374	0,05111
117	85	5,23077	-2,20541	1,79459	0,46983	0,32436	0,39513	0,18596	0,00575	0,00397	0,06895	0,04798
118	86	5,76923	-0,22000	3,78000	0,51819	0,68320	0,45030	0,61831	0,00634	0,00835	0,07604	0,10107
119	87	5,66667	-0,77000	3,23000	0,50899	0,58380	0,43980	0,49854	0,00623	0,00714	0,07469	0,08636
120	88	6,07692	-1,30810	2,69190	0,54583	0,48654	0,48184	0,38136	0,00668	0,00595	0,08010	0,07197
122	89	6,15385	-1,06371	2,93629	0,55274	0,53071	0,48972	0,43458	0,00676	0,00649	0,08111	0,07851
124	90	5,58333	-2,06047	1,93953	0,50150	0,35055	0,43125	0,21752	0,00614	0,00429	0,07359	0,05186
125	91	6,15385	-1,01104	2,98896	0,55274	0,54023	0,48972	0,44605	0,00676	0,00661	0,08111	0,07992
126	92	7,76923	-0,51000	3,49000	0,69783	0,63079	0,65526	0,55516	0,00854	0,00771	0,10240	0,09331
127	93	6,83333	-1,32932	2,67068	0,61377	0,48270	0,55935	0,37674	0,00751	0,00590	0,09007	0,07141
128	94	6,86667	-0,73754	3,26246	0,61677	0,58966	0,56277	0,50561	0,00755	0,00721	0,09051	0,08723
130	95	6,00000	-0,75630	3,24370	0,53892	0,58627	0,47396	0,50152	0,00659	0,00717	0,07909	0,08673
132	96	8,33333	-0,60789	3,39211	0,74850	0,61310	0,71306	0,53384	0,00916	0,00750	0,10984	0,09070
133	97	5,31250	-1,24000	2,76000	0,47717	0,49885	0,40350	0,39619	0,00584	0,00610	0,07002	0,07380
137	98	3,50000	-0,75055	3,24945	0,31437	0,58731	0,21776	0,50278	0,00385	0,00718	0,04613	0,08688
138	99	8,31250	-0,33179	3,66821	0,74663	0,66300	0,71093	0,59397	0,00914	0,00811	0,10957	0,09808
139	100	9,50000	-0,25061	3,74939	0,85330	0,67767	0,83262	0,61165	0,01044	0,00829	0,12522	0,10025
140	101	6,41667	-0,86889	3,13111	0,57635	0,56592	0,51666	0,47700	0,00705	0,00692	0,08458	0,08372
141	102	6,30769	-2,66441	1,33559	0,56656	0,24140	0,50549	0,08600	0,00693	0,00295	0,08314	0,03571
142	103	5,71429	-1,46932	2,53068	0,51326	0,45740	0,44468	0,34625	0,00628	0,00559	0,07532	0,06766
146	104	5,16667	-2,37009	1,62991	0,46408	0,29459	0,38856	0,15009	0,00568	0,00360	0,06810	0,04358
147	105	7,28571	-1,47435	2,52565	0,65441	0,45649	0,60571	0,34516	0,00801	0,00558	0,09603	0,06753
149	106	6,50000	-1,66489	2,33511	0,58383	0,42205	0,52519	0,30366	0,00714	0,00516	0,08568	0,06243
150	107	8,00000	-0,39195	3,60805	0,71857	0,65213	0,67891	0,58087	0,00879	0,00797	0,10545	0,09647
152	108	7,36364	-0,81380	3,18620	0,66140	0,57588	0,61369	0,48900	0,00809	0,00704	0,09706	0,08519
154	109	10,76923	-0,90869	3,09131	0,96730	0,55873	0,96269	0,46834	0,01184	0,00683	0,14195	0,08265
155	110	6,46154	1,53276	5,53276	0,58038	1,00000	0,52125	1,00000	0,00710	0,01223	0,08517	0,14793
156	111	7,06667	0,48644	4,48644	0,63474	0,81089	0,58327	0,77215	0,00777	0,00991	0,09315	0,11996
158	112	10,75000	-1,58662	2,41338	0,96557	0,43620	0,96072	0,32071	0,01181	0,00533	0,14169	0,06453
160	113	11,13333	0,68201	4,68201	1,00000	0,84623	1,00000	0,81474	0,01224	0,01035	0,14675	0,12518
162	114	6,06667	-0,56078	3,43922	0,54491	0,62161	0,48079	0,54410	0,00667	0,00760	0,07996	0,09196
164	115	7,36364	-1,22413	2,77587	0,66140	0,50172	0,61369	0,39965	0,00809	0,00613	0,09706	0,07422
165	116	8,00000	-2,23206	1,76794	0,71857	0,31954	0,67891	0,18015	0,00879	0,00391	0,10545	0,04727
166	117	9,06250	-0,10891	3,89109	0,81400	0,70328	0,78779	0,64250	0,00996	0,00860	0,11945	0,10404
167	118	6,84615	-1,11284	2,88716	0,61493	0,52183	0,56067	0,42388	0,00752	0,00638	0,09024	0,07720
171	119	7,37500	-0,76000	3,24000	0,66243	0,58560	0,61486	0,50072	0,00811	0,00716	0,09721	0,08663
172	120	4,53846	-0,39656	3,60344	0,40765	0,65129	0,32419	0,57986	0,00499	0,00796	0,05982	0,09635
174	121	6,70000	-0,88000	3,12000	0,60180	0,56391	0,54569	0,47459	0,00736	0,00690	0,08831	0,08342
175	122	6,33333	-1,81653	2,18347	0,56886	0,39464	0,50811	0,27064	0,00696	0,00483	0,08348	0,05838
177	123	9,53846	-0,97921	3,02079	0,85675	0,54598	0,83657	0,45298	0,01048	0,00668	0,12573	0,08077
179	124	7,14286	0,41000	4,41000	0,64158	0,79707	0,59108	0,75550	0,00785	0,00975	0,09415	0,11791
180	125	10,21429	-0,01484	3,98516	0,91745	0,72028	0,90582	0,66299	0,01123	0,00881	0,13463	0,10655
181	126	5,71429	-2,05781	1,94219	0,51326	0,35103	0,44468	0,21810	0,00628	0,00429	0,07532	0,05193
183	127	6,72727	-1,22485	2,77515	0,60425	0,50159	0,54849	0,39949	0,00739	0,00613	0,08867	0,07420

i	k	$\mathbf{D_{ik}}$	$\mathbf{B_{ik}}$	$4+\mathbf{B_{ik}}$	D_k^1	B_k^1	D_k^2	B_k^2	D_k^3	B_k^3	D_k^4	B_k^4
184	128	6,70000	-1,64533	2,35467	0,60180	0,42559	0,54569	0,30792	0,00736	0,00520	0,08831	0,06296
185	129	6,14286	-1,72544	2,27456	0,55176	0,41111	0,48860	0,29048	0,00675	0,00503	0,08097	0,06082
186	130	5,27273	-0,65567	3,34433	0,47360	0,60446	0,39942	0,52344	0,00580	0,00739	0,06950	0,08942
187	131	7,13333	0,80315	4,80315	0,64072	0,86813	0,59009	0,84112	0,00784	0,01061	0,09402	0,12842
190	132	4,63636	-1,18245	2,81755	0,41644	0,50925	0,33422	0,40872	0,00510	0,00623	0,06111	0,07533
192	133	9,00000	-1,30434	2,69566	0,80839	0,48722	0,78139	0,38218	0,00989	0,00596	0,11863	0,07207
193	134	5,76923	-1,12222	2,87778	0,51819	0,52013	0,45030	0,42184	0,00634	0,00636	0,07604	0,07694
194	135	8,92308	-0,84564	3,15436	0,80148	0,57012	0,77351	0,48207	0,00981	0,00697	0,11761	0,08434
195	136	4,88889	-1,73247	2,26753	0,43912	0,40984	0,36009	0,28895	0,00537	0,00501	0,06444	0,06063
196	137	8,75000	-0,86809	3,13191	0,78593	0,56607	0,75577	0,47718	0,00962	0,00692	0,11533	0,08374
197	138	5,86667	-1,02465	2,97535	0,52695	0,53777	0,46030	0,44308	0,00645	0,00658	0,07733	0,07955
198	139	4,85714	-0,36432	3,63568	0,43627	0,65712	0,35683	0,58688	0,00534	0,00803	0,06402	0,09721
203	140	10,45455	-0,11007	3,88993	0,93903	0,70307	0,93044	0,64225	0,01149	0,00860	0,13780	0,10401
206	141	5,36364	-1,44000	2,56000	0,48176	0,46270	0,40874	0,35264	0,00589	0,00566	0,07070	0,06845
210	142	6,30000	-0,39210	3,60790	0,56587	0,65210	0,50470	0,58083	0,00692	0,00797	0,08304	0,09647
212	143	8,85714	-0,26839	3,73161	0,79555	0,67446	0,76674	0,60777	0,00973	0,00825	0,11674	0,09977
214	144	8,45455	-0,46392	3,53608	0,75939	0,63912	0,72549	0,56519	0,00929	0,00781	0,11144	0,09455
215	145	7,66667	-1,15037	2,84963	0,68863	0,51505	0,64475	0,41571	0,00843	0,00630	0,10105	0,07619
216	146	7,66667	-0,82435	3,17565	0,68863	0,57397	0,64475	0,48670	0,00843	0,00702	0,10105	0,08491
217	147	5,71429	-0,35454	3,64546	0,51326	0,65889	0,44468	0,58901	0,00628	0,00806	0,07532	0,09747
219	148	8,33333	0,45404	4,45404	0,74850	0,80503	0,71306	0,76509	0,00916	0,00984	0,10984	0,11909
220	149	5,58333	0,17955	4,17955	0,50150	0,75542	0,43125	0,70532	0,00614	0,00924	0,07359	0,11175
221	150	4,64706	0,17256	4,17256	0,41741	0,75416	0,33531	0,70380	0,00511	0,00922	0,06125	0,11156
222	151	7,86667	0,37878	4,37878	0,70659	0,79143	0,66525	0,74870	0,00865	0,00968	0,10369	0,11708
223	152	8,00000	-0,39656	3,60344	0,71857	0,65129	0,67891	0,57986	0,00879	0,00796	0,10545	0,09635
224	153	8,61538	0,73551	4,73551	0,77384	0,85590	0,74197	0,82639	0,00947	0,01047	0,11356	0,12662
228	154	6,35714	-0,79300	3,20700	0,57100	0,57964	0,51055	0,49353	0,00699	0,00709	0,08379	0,08575
229	155	7,58333	0,51670	4,51670	0,68114	0,81636	0,63621	0,77874	0,00833	0,00998	0,09995	0,12076
230	156	7,54545	0,25099	4,25099	0,67774	0,76833	0,63233	0,72087	0,00829	0,00939	0,09946	0,11366
231	157	8,61538	-0,71877	3,28123	0,77384	0,59305	0,74197	0,50969	0,00947	0,00725	0,11356	0,08773
241	158	6,58696	-0,37162	3,62838	0,59165	0,65580	0,53411	0,58529	0,00724	0,00802	0,08682	0,09701
244	159	9,38462	1,07330	5,07330	0,84293	0,91696	0,82080	0,89994	0,01031	0,01121	0,12370	0,13565
245	160	6,75000	-0,10925	3,89075	0,60629	0,70322	0,55081	0,64243	0,00742	0,00860	0,08897	0,10403
246	161	6,00000	0,06595	4,06595	0,53892	0,73489	0,47396	0,68058	0,00659	0,00899	0,07909	0,10871
247	162	6,71429	-0,59241	3,40759	0,60308	0,61589	0,54715	0,53721	0,00738	0,00753	0,08850	0,09111
251	163	2,91667	0,30451	4,30451	0,26198	0,77800	0,15799	0,73253	0,00321	0,00951	0,03844	0,11509

Tabla 1.- Valores de las muestras originales y sus respectivas transformaciones a una métrica común

2 Resultados de la prueba de Wilcoxon y del T-test

En este apartado se muestran, primeramente, los resultados calculados con SPSS v.17 de la aplicación de la prueba de la suma de rangos con signo de Wilcoxon a cada una de los pares de muestras transformadas y, seguidamente, de la aplicación del T-test para la comparación de medias entre las misma muestras pareadas.

```
GET
  FILE='C:\Documents and Settings\Rosa\Escritorio\20091130-DvsB.sav'.
NPAR TESTS
  /WILCOXON=DkP1 DkP2 DkP3 DkP4 WITH BkP1 BkP2 BkP3 BkP4 (PAIRED)
  /STATISTICS DESCRIPTIVES
  /MISSING ANALYSIS.
```

Pruebas no paramétricas

```
[Conjunto_de_datos1] C:\Documents and Settings\Rosa\Escritorio\20091130-Dv
sB.sav
```

Estadísticos descriptivos

	N	Media	Desviación típica	Mínimo	Máximo
DkP1	163	,501418698	,1835073606	,1235034	1,0000000
DkP2	163	,431165755	,2093645919	,0000000	1,0000000
DkP3	163	,006134969	,0022452533	,0015111	,0122352
DkP4	163	,073581533	,0269290973	,0181237	,1467467
BkP1	163	,501771804	,1695442488	,1700151	1,0000000
BkP2	163	,399714166	,2042738884	,0000000	1,0000000
BkP3	163	,006134969	,0020729518	,0020787	,0122266
BkP4	163	,074227860	,0250809364	,0251506	,1479315

Prueba de los rangos con signo de Wilcoxon

Rangos

		N	Rango promedio	Suma de rangos
BkP1 - DkP1	Rangos negativos	86[a]	79,97	6877,00
	Rangos positivos	77[b]	84,27	6489,00
	Empates	0[c]		
	Total	163		
BkP2 - DkP2	Rangos negativos	98[d]	82,31	8066,00
	Rangos positivos	65[e]	81,54	5300,00
	Empates	0[f]		
	Total	163		
BkP3 - DkP3	Rangos negativos	86[g]	80,16	6894,00
	Rangos positivos	77[h]	84,05	6472,00
	Empates	0[i]		
	Total	163		
BkP4 - DkP4	Rangos negativos	85[l]	78,67	6687,00
	Rangos positivos	78[k]	85,63	6679,00
	Empates	0[l]		
	Total	163		

a. BkP1 < DkP1
b. BkP1 > DkP1
c. BkP1 = DkP1
d. BkP2 < DkP2
e. BkP2 > DkP2
f. BkP2 = DkP2
g. BkP3 < DkP3
h. BkP3 > DkP3
i. BkP3 = DkP3
j. BkP4 < DkP4
k. BkP4 > DkP4
l. BkP4 = DkP4

Estadísticos de contraste[b]

	BkP1 - DkP1	BkP2 - DkP2	BkP3 - DkP3	BkP4 - DkP4
Z	-,321[a]	-2,292[a]	-,350[a]	-,007[a]
Sig. asintót. (bilateral)	,748	,022	,727	,995

a. Basado en los rangos positivos.

b. Prueba de los rangos con signo de Wilcoxon

```
T-TEST PAIRS=DkP1 DkP2 DkP3 DkP4 WITH BkP1 BkP2 BkP3 BkP4 (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.
```

Prueba T

```
[Conjunto_de_datos1] C:\Documents and Settings\Rosa\Escritorio\20091130-Dv
sB.sav
```

Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	DkP1	,501418698	163	,1835073606	,0143734058
	BkP1	,501771804	163	,1695442488	,0132797305
Par 2	DkP2	,431165755	163	,2093645919	,0163987004
	BkP2	,399714166	163	,2042738884	,0159999658
Par 3	DkP3	,006134969	163	,0022452533	,0001758618
	BkP3	,006134969	163	,0020729518	,0001623661
Par 4	DkP4	,073581533	163	,0269290973	,0021092497
	BkP4	,074227860	163	,0250809364	,0019644906

Correlaciones de muestras relacionadas

		N	Correlación	Sig.
Par 1	DkP1 y BkP1	163	,532	,000
Par 2	DkP2 y BkP2	163	,532	,000
Par 3	DkP3 y BkP3	163	,532	,000
Par 4	DkP4 y BkP4	163	,532	,000

Prueba de muestras relacionadas

		Diferencias relacionadas		
		Media	Desviación típ.	Error típ. de la media
Par 1	DkP1 - BkP1	-,0003531067	,1711342478	,0134042688
Par 2	DkP2 - BkP2	,0314515890	,2000400492	,0156683459
Par 3	DkP3 - BkP3	,0000000000	,0020932533	,0001639563
Par 4	DkP4 - BkP4	-,0006463270	,0251983603	,0019736879

Prueba de muestras relacionadas

		Diferencias relacionadas				
		95% Intervalo de confianza para la diferencia				
		Inferior	Superior	t	gl	Sig. (bilateral)
Par 1	DkP1 - BkP1	-,0268227279	,0261165145	-,026	162	,979
Par 2	DkP2 - BkP2	,0005110590	,0623921189	2,007	162	,046
Par 3	DkP3 - BkP3	-,0003237670	,0003237669	,000	162	1,000
Par 4	DkP4 - BkP4	-,0045437996	,0032511457	-,327	162	,744

3 CE-PE2: Extrapolaciones de costes

Las extrapolaciones de los costes temporales y económicos de la calibración CE para otros tamaños de bancos de ítems empleando solo aportaciones recabadas con pruebas de campo PE2, se realizan a partir de los valores y tasas identificados durante la calibración CE-PE2 realizada. Dichos valores se han recogido la Tabla 2 y la Tabla 3.

Para elaborar estimaciones de costes de calibraciones con tamaños de bancos alternativos se conservarán de la calibración original (con n=252) los apartados específicos enunciados en la Tabla 2 y que son: el número de ítems por cuestionario, las valoraciones mínimas a disponer a la hora de realizar la estimación de los parámetros, la tasa de abandono de expertos (abandono de participación sin completar de cuestionario), tasa de descarte de experto (participación inadecuada por parte del experto), promedio de minutos telefónicos por cuestionario comprometido, tiempo medio empleado por el experto para completar un cuestionario, número medio de emails enviados por cuestionario comprometido y tiempo invertido por cada revisor para supervisar la mitad del banco original.

Ítems por cuestionario	42
Valoraciones por ítem	7
Tasa abandono experto	48,1%
Tasa cumplimiento (cuestionario completado por expertos)	51,9%
Tasa descarte experto	4,3%
Tasa cuestionario validado (experto serio)	95,7%
T. duración llamadas por cuestionario comprometido (h)	0,10
T. completado de cuestionario por experto	0,83
Emails envíados por cuestionario comprometido	1,2
T. supervisando n/2 ítems por supervisor	3

Tabla 2.- Valores para extrapolación de costes heredados de la CE-PE2

A partir de estos valores y por cada tamaño de banco de ítems, se determina (Tabla 3): el número de cuestionarios distintos a construir, el número de cuestionarios completados a recabar (considerando la tasa de descarte de expertos) y el número de expertos a involucrar (considerando la tasa de incumplimiento). Para una mejor comprensión, seguidamente se ilustra con un ejemplo la interpretación de los apartados mencionados. Si quisiéramos calibrar un banco con 500 ítems, habría que confeccionar 12 cuestionarios distintos e involucrar a 170 expertos para recabar 88 cuestionarios completados, de los cuales, y si se conserva la tasa de descarte, finalmente se emplearán las valoraciones de 84 de los 88 cuestionarios.

	0*5 n	n	2*n	3*n	4*n
Factor ajuste por Tamaño BI	0,5	1	2	3	4
Factor corrector entregables	0,75	1	1,25	1,5	1,75
Tamaño BI (del banco de ítems)	125	250	500	750	1000
N. Cuestionarios	3	6	12	18	24
Cuestionarios completados (7 val./ítem)	21	42	84	126	168
Cuestionarios extra por descarte de experto	22	44	88	132	176
N. participantes pasivos a captar	43	85	170	255	340

Tabla 3.- Volumen de expertos a captar y de cuestionarios completados a recabar

Para estimar los tiempos de los participantes activos, se han tenido en cuenta la existencia de costes fijos y costes variables. Los *costes fijos* son apartados cuyos valores son independientes del volumen del banco de ítems a calibrar, y por ello, sus valores son invariables en todas las estimaciones, como por ejemplo, el tiempo invertido en *formación*. Los *costes variables* son apartados cuyos valores aumentan conforme crecen los tamaños de los bancos de ítems a calibrar, ejemplo de ello son los entregables y los tiempos dedicados al análisis de los datos. Así, mientras que los costes fijos se han conservado, los valores de los costes variables se han calculado por aplicación dos factores correctores con respecto los costes variables del banco original. En la primera línea en la Tabla 3 están los valores correctores concretos aplicados a los *entregables* y a aspectos de *diseño* y en la segunda línea los aplicados a aspectos de *planificación-gestión* y de *implementación*. En la Tabla 4 se muestran, desglosados por los apartados considerados, los valores de los costes estimados para realizar calibraciones CE con distintos tamaños de bancos.

		CE-PE2 (7 valoraciones)				
	Apartados	**0,5*n**	**n**	**2*n**	**3*n**	**4*n**
Fase 1: **Recogida de**	Formación	128	128	128	128	128
	Planificación y gestión	81,3	95	121,3	147,5	173
	T. pasivos	25,8	51,7	103,3	155	206,7
	Implementación	139,3	248,4	466,9	685,3	903,8
	Entregables	123,8	155	186,3	217,5	248,8
	Subtotal(h)	***498,2***	***678,1***	***1005,8***	***1333,3***	***1660,3***
Fase 2: **Análisis y calibración**	Formación	90	90	90	90	90
	Planificación y gestión	45	55	65	75	85
	Análisis resultados	9	12	15	18	21
	Implementación	31,5	42	52,5	63	73,5
	Entregables	67,5	90	112,5	135	157,5
	Subtotal(h)	***243***	***289***	***335***	***381***	***427***
	Coste temporal(h)	**741,2**	**967,1**	**1340,8**	**1714,3**	**2087,3**

Tabla 4.- Extrapolación de costes para realizar calibraciones CE con bancos de distintos tamaños

4 CT-PT2: Extrapolaciones de costes

Las extrapolaciones de los costes temporales y económicos de la calibración CT para otros tamaños de bancos de ítems empleando solo aportaciones recabadas con pruebas de campo PT2, se realizan a partir de los valores y tasas identificados durante la calibración CT-PT2 realizada. Dichos valores se han recogido la Tabla 5 y la Tabla 6.

Para elaborar estimaciones de costes de calibraciones con tamaños de bancos alternativos se conservarán de la calibración original (con n=250) los apartados específicos enunciados en la Tabla 5 y que son: el número de ítems total y de anclaje por cuestionario, las valoraciones mínimas a disponer a la hora de realizar la estimación de los parámetros, la tasa de no validación (por aplicación de criterios de descarte y de no validación), número de llamadas por cuestionario completado, promedio de alumnos por centro administrados, promedio de puestos de ordenador disponibles en los laboratorios, diversos valores relacionados con los kilómetros desplazados por los participantes activos hasta los centros administrados, tasa de descarte de experto, promedio del tiempo invertido por un sujeto anónimo completando el cuestionario electrónico, y tiempo invertido por cada revisor para supervisar la mitad del banco original.

Ítems por cuestionario	60
Ítems de anclaje por cuestionario	22
Vol. ítems anclaje por cuestionario (20%-40%)	36,67%
Valoraciones por ítem	500
Tasa no validación (descartes e invalidados)	3,20%
T. llamadas por cuestionario rellenado (min)	0,0045
N. medio de alumnos (cuestionarios) por centro	235
Tamaño medio de laboratorios	20
km desplazados por cuestionarios rellenado	0,66
Velocidad media de desplazamiento	60
N. viajeros (p. activos)	2,6
T. completado cuestionario por sujeto de pt4	0,315167
T. supervisando n/2 ítems por supervisor	3

Tabla 5.- Valores para extrapolación de costes heredados de la CT-PT2

Nuevamente, a partir de estos valores y por cada tamaño de banco de ítems, se determina (Tabla 6): los mismos factores correctores que los concretados para las extrapolaciones de la calibración CE, los mismos tamaños de bancos de ítems, el número de cuestionarios distintos a construir, por cada cuestionarios el número de ítems de anclaje y el resto (hasta 60) de ítems propios del cuestionarios, el número de cuestionarios completados a recabar (suponiendo nuevamente que la tasa de abandono en los laboratorios PT4 es 0) y el número de sujetos anónimos a involucrar (considerando la tasa de no validación).

	0*5 n	n	2*n	3*n	4*n
Factor Tamaño BI	0,5	1	2	3	4
Factor corrector entregables	0,75	1	1,25	1,5	1,75
Tamaño banco de ítems (BI)	125	250	500	750	1000
N. Cuestionarios	3	6	12	18	24
Ítems anclaje	27	22	20	19	19
Ítems propios del cuestionario	33	38	40	41	41
Valoraciones mínimas	1500	3000	6000	9000	12000
N. participantes pasivos a captar	1548	3096	6192	9288	12384

Tabla 6.- Volumen de expertos a captar y de cuestionarios completados a recabar

Para estimar los tiempos de los participantes activos, se han aplicado a los costes variables los mismos factores correctores que los considerados en la calibración CT a los mismos aspectos. En la Tabla 7 se han recogido las estimaciones resultantes para realizar calibraciones CT con distintos tamaños de bancos y desglosadas por los apartados habituales.

		CT-PT2 (500 aportaciones)				
	Apartados	**0,5*n**	**n**	**2*n**	**3*n**	**4*n**
Fase 1: **Recogida de Datos**	Formación	158	158	158	158	158
	Planificación y gestión	92,5	95	100	105	110
	T. pasivos	495,4	990,8	1981,5	2972,3	3963
	Implementación	265,6	428,7	754,8	1081	1407,2
	Entregables	157,5	200	242,5	285	327,5
	Subtotal(h)	***1169***	***1872,5***	***3236,8***	***4601,3***	***5965,7***
Fase 2: **Análisis y calibración**	Formación	80	80	80	80	80
	Planificación y gestión	63,5	66	68,5	71	73,5
	Análisis resultados	6	8	10	12	14
	Implementación	20,3	27	33,8	40,6	47,3
	Entregables	60	80	100	120	140
	Subtotal(h)	***229,8***	261	***292,3***	***323,6***	***354,8***
	Coste temporal(h)	**1398,8**	***2133,5***	**3529,1**	**4924,9**	**6320,5**

Tabla 7.- Extrapolación de costes para realizar calibraciones CE con bancos de distintos tamaños

ANEXO7 Póster BPMN 1.1

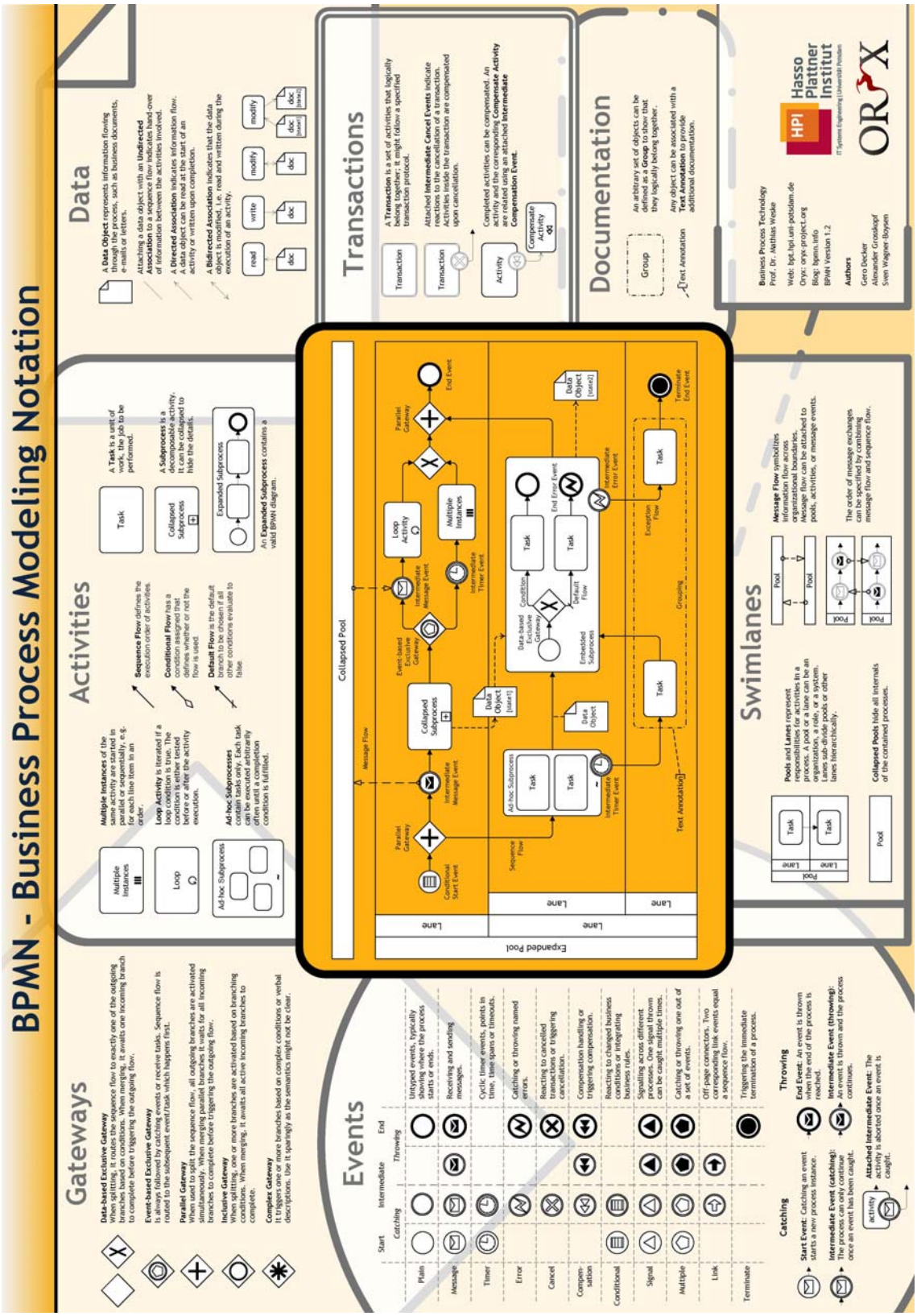
Un Procesos de Negocio es una colección de actividades que tomando una o varias clases de entradas crean una salida que tiene valor para un cliente (Hammer y Champy, 1993). Los procesos de negocio se pueden representar gráficamente empleando diversas notaciones.

La BPMN (Business Process Managment Notation) es una notación estándar para el modelado de flujos de procesos de negocio y servicios web. En concreto, BPMN define la notación y la semántica de un Diagrama de Procesos de Negocio.

La primera versión de la notación fue desarrollada por BPMI (Business Process Managment Initiative) en 2004. Tras la fusión con OMG (Object Manangmet Group) en 2005, la versión BPMN 2.0 salió a la luz en 2006 (http://bpmn.org/).

Este anexo recoge sucintamente (1) los elementos centrales y (2) la lista completa de elementos que se pueden emplear a la hora de definir un diagrama de proceso de negocio.

Las personas interesadas en ampliar conocimientos pueden consultar, entre otros, el manual de referencia de la notación (BPMI.org, 2004) y el informe blanco (White, 2004) donde se confrontan la notación BMPN y los diagramas de actividad de UML.


BPMN - Business Process Modeling Notation
Gateways
Activities
Data
Transactions
Documentation
Events
Swimlanes
Collapsed Pool
Expanded Pool
Business Process Technology
Prof. Dr. Mathias Weske
Web: bpt.hpi.uni-potsdam.de
Oryx: oryx-project.org
Blog: bpmn.info
BPMN Version 1.2
Authors
Gero Decker
Alexander Grosskopf
Sven Wagner-Boysen
Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam
ORYX

Referencias bibliográficas

Aguilera, J. M. y M. J. González (2008). Test oposiciones cuerpo de tramitación procesal y administrativa de la administración de justicia. Turno libre, Ed. Cep.

Andersen, E. B. (1970). "Asymptotic properties of conditional maximum likelihood estimators." Journal of the Royal Society, Series B **32**: 283-301.

Angoff, W. H. (1984). Scales, norms and equivalent scores. New Jersey (USA), Educational Testing Service. Princeton.

Armendáriz, A. J., M. Izquierdo, A. Tapias, J. López-Cuadrado, J. Á. Vadillo, T. A. Pérez y J. Gutiérrez (2002). HEUSKLEARNING: An educational hyperenvironment to learn the Basque language. En A. M. Vilas, J. A. M. González y I. S. d. Zaldívar (eds.) Proc. of Educational technology - Information society and education: monitoring a revolution, Badajoz (España), Junta de Extremadura (CECT), **2**: 689-694.

Armendáriz, A. J., J. López-Cuadrado, J. Á. Vadillo y T. A. Pérez (2004). HEUSKLEARNING: Un hiperentorno educativo para el aprendizaje del euskara. En (eds.) Proc. of IV Encuentro Europa-América Latina sobre Formación y Cooperación Tecnológica y Profesional, Isla de Margarita (Venezuela):

Arruabarrena, R. (2005). Filtrado de un banco de ítems. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 02-2005): 60.

Arruabarrena, R. y A. J. Armendariz (2008). Estimación de los parámetros de los ítems de un sistema de e-learning vía expertos. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 02-2008): 244.

Arruabarrena, R., J. López-Cuadrado y A. J. Armendariz (2007). Consideraciones para el cómputo de costes de calibraciones de bancos de ítems. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 06-2007): 37.

Arruabarrena, R., J. López-Cuadrado y A. J. Armendariz (2010). Calibración 3PL-TRI de ítems: procesos BPM, ejecución, análisis y mejora. Una investigación empírica. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 09-2010): 44.

Arruabarrena, R. y T. A. Pérez (2005a). Pruebas de campo para calibrar un banco de ítems vía expertos. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 08-2005): 96.

Arruabarrena, R. y T. A. Pérez (2005b). Una experiencia arbitrando incidencias producidas en pruebas de campo. En M. O. Cantero (eds.) Proc. of VI congreso nacional de Informática Educativa. I Simposio Nacional de Tecnologías de la Información y las Comunicaciones en la Educacion: SIntice-2005 (SINTICE-CEDI'05), Granada, Thomson Paraninfo (Spain): 161-166.

Arruabarrena, R. y T. A. Pérez (2010). Calibración de ítems con expertos: procesos BPM, ejecución, análisis y mejora. Una investigación empírica. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 01-2010): 53.

Arruabarrena, R., T. A. Pérez, J. Gutiérrez, J. López-Cuadrado y J. A. Vadillo (2001). Compendio de técnicas para evaluación de sistemas hipermedia adaptativos. En (eds.) Proc. of Simposium Internacional de Informática Educativa, SIIE, Instituto Superior Politécnico de Viseu, Viseu, Portugal, Escola Superior de Educaciçao: 10.

Arruabarrena, R., T. A. Pérez, J. Gutiérrez, J. López-Cuadrado y J. A. Vadillo (2002). On Evaluating Adaptive Systems for Education. En P. D. Bra, P. Brusilovsky y R. Conejo (eds.) Proc. of AH2002, 2nd. International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Málaga, Lecture Notes in Computer Science, Springer-Verlag, **2347**: 363-367.

Arruabarrena, R., S. Sanz-Santamaría y J. Gutiérrez (2007). Desarrollo eficiente de pruebas de campo. En I. F. d. Castro (eds.) Proc. of Simposio Nacional de Tecnologías de la Información y las Comunicaciones en la Educación: Sintice-2007, incluido en el II Congreso Español De Informática: CEDI'07 (SINTICE-CEDI'07), Zaragoza (España), Nuevos retos científicos y tecnológicos en Ingeniería Informática, Thomsom, **1**: 309-312.

ASC (1997). User's manual for the XCALIBRE marginal maximum-likelihood estimation program (second edition). St. Paul, Minnesota (USA), Assessment Systems Corporation.

Baker, F. B. (1992). Item response theory: parameter estimation techniques. New York (USA), Marcel Dekker.

Barba-Romero, S. y J. C. Pomerol (1997). Decisiones multicriterio. Fundamentos teóricos y utilización práctica, Servicio de Publicaciones de la Universidad de Alcalá.

Barbero, M. I. (1996). Chapter "Banco de ítems." Psicometría. J. Muñiz. Madrid (España), Editorial Universitas, S.A.: 139-170.

Basili, V. R. (1985). Quantitative Evaluation of Software Engineering Methodology". En (eds.) Proc. of First Pan Pacific Computer Conference, Melbourne (australia), **vol.1**: 379-398.

Basili, V. R. (1999). "Building knowledge through families of experiments." IEEE Transactions of Software Engineering **25**(4): 456-473.

Basili, V. R., R. W. Selby y D. H. Hutchens (1986). "Experimentation in Software Engineering." IEEE Transactions of Software Engineering **12**(7): 733-743.

Binet, A. y T. Simon (1905). "Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux." L'Année psychologique **11**: 191-244.

Birnbaum, A. (1968). Chapter "Some latent trait models and their use in inferring an examinee's ability." Statistical theories of mental test scores. F. M. Lord y M. R. Novick. Reading (USA), Addison-Wesley: chapters 17-20.

Bock, R. D. y M. Lieberman (1970). "Fitting a response model for n dichotomously scored items." J. Psychometrika **35**: 179-197.

BPMI.org (2004). Business Process Modeling Notation (BMPN), version 1.0: 296.

Brusilovsky, P. y L. Pesin (1998). "Adaptive Navigation Support in Educational Hypermedia: An Evaluation of the ISIS-Tutor." Journal of Computing and Information Technology (CIT) **6**(1): 27-38.

Bunderson, C. V., D. K. Inouye y J. B. Olsen (1989). Chapter "The four generations of computerized educational measurement." Educational Measurement. R. L. Linn. New York: MacMillan: (3rd ed.) 367-407.

Burke, N. W., B. D. Kaufman y N. L. Webb (1985). The Wisconsin item bank: development, operation and related issues Madison Wisconsin Department of Public Instruction:

Calvi, L. (2000). Formative Evaluation of Adaptive CALLware: A Case Study. En P. Brusilovsky, O. Stock y C. Strapparava (eds.) Proc. of International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems: AH2000, Trento (Italy), Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg, **1892**: 276-279.

Campbell, D. T. y J. C. Stanley (1973). Experimental and Quasi-Experimental Design for Research. 1st edition. Chicago (EE.UU), Rand McNally, 1966. (traducido por Kitaigorodzki, M., Diseños Experimentales y Cuasi-Experimentales en la Investigación Social, Buenos Aires (Argentina): Ed. Amorrortu, 1973).

Caro, J. L. (1988). Eficacia de las pruebas objetivas para la enseñanza de las técnicas de expresión gráfica en la ingeniería. Expresión Gráfica y Proyectos de Ingeniería. Bilbao, UPV-EHU**:** 383.

Cawsey, A. J., R. B. Jones y J. Pearson (2000). "The Evaluation of a personalised Health Information System for Patients with Cancer." User Modeling and User-Adapted Interaction **10**: 47-72.

Cloquell, V., M. C. Santamarina, M. García-Melón y M. A. Sánchez (2001). A new procedure for the numerical values normalization in multicriteria decision techniques. En (eds.) Proc. of Proceedings of the 54th Conference of the European Multicriteria Working Group, Durbuy (Bélgica):

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales." Educational and Psychological Measurement **20**: 37-46.

Cohen, J. (1992). "A power primer." Psychological Bulletin **112**: 155-159.

Conejo, R., E. Guzmám, E. Millán, M. Trella, J. L. Pérez-De-La-Cruz y A. Ríos (2004). "SIETTE: A Web-Based Tool for Adaptive Testing." International Journal of Artificial Intelligence in Education **14**: 1-33.

Conejo, R., E. Guzmám y J. L. Pérez (2008). Un Estudio sobre la Dificultad de los Ítems en Tests de Informática. En R. Peña, R. Castillo y M.Anguita (eds.) Proc. of XIV Jornadas de Enseñanza Universitaria de la Informatica, Granada (España): 241-248.

Cook, T. D. y D. T. Campbell (1979). Quasi-Experimentation - Design and Analysis Issues for Field Settings. Boston, MA (EE.UU), Houghton Mifflin Company:

Cornish, G. y R. Wines (1977). Mathematics Profile Series. Hawthorn, Victoria, Australian Council for Educational Research.

Council_of_Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment, Cambridge University Press.

Cronbach, L. J. (1951). "Coefficient alpha and the internal structure of tests." J. Psychometrika **16**: 297-334.

Cuesta, M. (1996). Chapter "Unidimensionalidad." Psicometría. J. Muñiz. Madrid (España), Editorial Universitas, S.A.: 239-292.

Chen, W.-H. y D. M. Thissen (1997). "Local dependence indices for item pairs using item response theory." J. Educational and Behavioral Statistics **22**: 265-289.

Chen, W.-H. y D. M. Thissen (1999). "Estimation of item parameters for the three-parameter logistic model using the marginal likelihood of summed scores." British Journal of Mathematical and Statistical Psychology **52**: 19-37.

Choppin, B. H. (1981). Chapter "Educational measurement and the item bank model." Issues in evaluation and accountability. C. Lacey y D. Lawton. London, Methuen & Co.

Dalkey, N. C., B. Brown y S. Cochran (1970). "The Delphi method, III. Use of self rating to improve group estimates." Technological Forecasting and Social Change **1**: 283-291.

Davidove, E. A. y R. A. Reiser (1991). "Comparative acceptability and effectiveness of teacher-revised and designer-revised instruction." Educational Technology Research and Development **39**(2): 29-38.

Dix, A., J. Finlay, G. Abowd y R. Beale (1998). Human-Computer Interaction, 2nd edition, Pearson Education Limited.

Dolado, J. J. y L. Fernández (2000). Medición para la getión en la Ingeniería del Software. Madrid (España), Ra-Ma.

Draper, S. W., M. I. Brown, F. P. Henderson y M. E. (1996). Integrative evaluation: an emerging role for classroom studies of CAL. En M. R. Kibby y J. R. Hartley (eds.) Proc. of Computer Assisted Learning, **26**: 17-32.

Eignor, D. R. (1985). An investigation of the feasibility and`practical outcomes of pre-equating the SAT verbal and mathematical sections. New Jersey (USA), Princeton, Educational Testing Service (RR-85-10):

Elliot, C. D. (1983). British ability scales. Manuals 1-4. Windsor, England, NFER-Nelson:

Ewing, J. (2000). e-Learning is not always easy learning. Case Study submitted to the International Online Tutoring Skills e-workshop 12 May 2000. In http://otis.scotcit.ac.uk/casestudy/ewing.doc.

Fenton, N. E. y S. L. Pfleeger (1998). Software Metrics: A Rigorous and Practical Approach, 2nd edition. Londres (Reino Unido), PWS.

Finney, K., K. Rennolls y A. Fedorec (1998). "Measuring the comprehensibility of Z specifications." Journal of Systems and Software **42**(1): 3-15.

Fleiss, J. L. (1971). "Measuring nominal scale agreement among many raters." Psychological Bulletin **76**: 378-381.

Fleiss, J. L. (1981). Statistical Methods for Rates and Proportions, 2nd edition. New York: John Wiley.

Fleiss, J. L. (1986). The design and analysis of clinical experiments. New York: John Wiley.

Fleiss, J. L. y J. Cohen (1973). "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability." Educational and Psychological Measurement **33**: 613-619.

Frechtling, J. y L. Sharp (1997). User-Friendly Handbook for Mixed Methods Evaluations Division of Research, Evaluation and Comunication. Nationale Science Foundation´s Directorate for Education and Human Resources (HER):

Fusaro, P., F. Lanubile y G. Visaggio (1997). "A replicated experiment to assess requeriments Inspection Techniques." Empirical Software Engineering **2**(1): 39-57.

Gagné, R. M. y L. J. Briggs (1979). Principles of Instructional Design (2nd edition). Orlando, New York (USA), Holt, Rinehart and Winston Inc.

Gagné, R. M., L. J. Briggs y W. W. Wager (1988). "Principles of instructional design."

García-Pérez, J., S. Cruz-Rambaud y C. García-García (2004). Chapter "Proceso iterativo de valoración en el método de las dos betas." Programación, selección, control y valoración de proyectos. IV reunión científica. R. Herrerías y J. Callejón. Granada, Universidad de Granada. Departamento de Métodos Cuantitativos para la Economía y la Empresa: 37-63.

García Cueto, E. (1996). Chapter "Software psicométrico." Psicometría. J. Muñiz. Madrid (España), Editorial Universitas, S.A.: 613-642.

Glas, C. A. W. (2000). Chapter " Item calibration and parameter drift." Computerized adaptive testing: theory and practice. W. J. v. d. Linden y C. A. W. Glas. Dordrecht (The Netherlands), Kluwer Academic Press: 183-199.

González, A., T. A. Pérez, R. López, J. A. Carro y J. Gutiérrez (1999). Hezinet: Una alternativa telemática al tradicional aprendizaje de idiomas. En (eds.) Proc. of Jornadas de transferencia tecnológica de inteligencia artificial, TTIA, Murcia, España, AEPIA, **II**: 1-9.

Green, N. y S. Carberry (1999). "A computational mechanism for initiative in answer generation." User Modeling and User-Adapted Interaction **9**(1-2): 93-132.

Gutiérrez, J., T. A. Pérez, P. Lopistéguy y I. Usandizaga (1995). Sistemas Tutores Inteligentes: una forma de conseguir Sistemas Hipermedia Educativos. En (eds.) Proc. of Conferencia de la Asociación Española para la Inteligencia Artificial, CAEPIA, Alicante, España, Asociación Española Para la Inteligencia Artificial, AEPIA: 249-258.

HABE, Ed. (1999). Helduen euskalduntzearen oinarrizko kurrikulua-(HEOK). Donostia/San Sebastián, Helduen Alfabetatze eta Berreuskalduntzerako Erakundea (HABE) - Eusko Jaurlaritza/Gobierno Vasco.

Hambleton, R. K. y H. Swaminathan (1985). Item response theory: principles and applications. Boston (USA), Kluwer-Nijhoff Publishing.

Hambleton, R. K., H. Swaminathan y H. J. Rogers (1991). Fundamentals of Item Response Theory. Newbury Park, CA, Sage.

Hambleton, R. K., J. N. Zaal y J. P. M. Pieters (1991). Chapter "Computerized adaptive testing: theory, applications, and standards." Advances in educational and psychological testing. R. K. Hambleton y J. N. Zaal. Norwell, Massachussets (USA), Kluwer Academic Publishers.

Harvey, J. (1998). Evaluation Cookbook. Heriot Watt University Edinburgh, Learning Technology Dissemination Initiative, Institute for Computer Based Learning.

Helmer, O. y N. Rescher (1959). "On the epistemology of the inexact sciences." Management Science **6**: 25-52.

Henning, G. (1986). Chapter "Item banking via dBase II: The UCLA ESL Proficiency Examination experience." Technology and language testing. C. W. Stansfield. Washington, DC (EE:UU), TESOL (Teachers of English to Speakers of Other Languages, Inc.): 69-77.

Herrerías, R., F. Palacios, J. Callejón y E. Pérez (1999). Un método para contrastar la bondad de un experto en la metodología de PERT. En (eds.) Proc. of Proc. de la XIII Reunión Anual de ASEPELT-España, Universidad de Burgos: 109-116.

Hill, P. W. (1985). The tests of reading comprehension (TORCH), Comunicación presentada en la reunión anual de la IEA, Oxford.

Hiscox, M. D. y E. Brzenzinski (1980). A guide to item banking in education (prefared for the Annual conference on large-scale assessment). Portland, OR, Northwest Regional Educational Laboratory, Assessment and Evaluation Division**:** 138.

Hontangas, P., V. Ponsoda, J. Olea y F. J. Abad (2000). "Los tests adaptativos informatizados en la frontera del siglo XXI: una revisión." J. Metodología de las Ciencias del Comportamiento **2**(2): 183-216.

Höök, K. (2000). "Steps to take before Intelligent User Interfaces become real." Journal of Interaction with Computers(12).

Höök, K. y M. Svensson (1999). Evaluating Adaptive Navigation Support. En (eds.) Proc. of Proceeding of IUI'99, Los Angles, CA, USA. (poster):

J.C.S.E.E. (1994). The program evaluation standards: how to assess evaluations of educational programs. 2nd edition. The Joint Committee on Standards for Educational Evaluation & Sanders, James R. Thousand Oaks, California (EE.UU), Sage Publications.

Jeffries, R., J. R. Miller, C. Wharton y K. M. Uyeda (1991). User interfaces evaluation in the real world: A comparison of four techniques. En (eds.) Proc. of ACM Humman Computer Interaction Conf.: 119-124.

Juristo, N. y A. M. Moreno (2001). Basics of software Engineering Experimentation, Kluwer Academinic Publishers.

Kachigan, S. K. (1986). Statistical Analysis. An Interdisciplinary Introduction Univariate & Multivariate Methods. New York, Radius Press.

Kaiser, H. F. (1974). "An index of factorial simplicity." Psychometrica **39**: 31-36.

Karat, C. M., R. Campbell y T. Fiegel (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. En (eds.) Proc. of Human Factors in Computing Systems Conf., New York, ACM: 397-404.

Kearns, S. P. (1998). Scoring, item analysis, reliability, and validity. En (eds.) Proc. of Workshop II Test Analysis, Charleston, SC (EEUU), Trident Techical College:

Kingsbury, G. G. y D. J. Weiss (1983). Chapter "A comparison of IRT-based Adaptive Mastery Testing and Sequential Mastery Testing Procedure." New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing. D. J. Weiss. New York, Academic Press: XXX- Revisar si es libro.

Kitchenham, B. A., L. Pickard y S. L. Pfleeger (1995). "Case Studies for Method and Tool Evaluation." IEEE Software(July 1995): 52-62.

Kolen, M. J. y R. L. Brennan (1995). Test equating: methods and practices. New York (USA), Springer-Verlag.

Kubeš, T. (2007). Application of Hypermedia Systems in e-Learning. Department of Computers. Faculty of Electrical Engineering. Praga (R. Chequia), Czech Technical University in Prague.**:** 193.

Kubinger, K. D. (1985). On a Rash model based test for noncomputerized adaptative testing. En (eds.) Proc. of 13th IPN Conference on Latent Trait and Latent Class Models in Educational Research, Kiel:

Kuder, G. B. y M. W. Richardson (1937). "The theory of estimation of test reliability." Psychometrica **2**: 151-160.

Lafourcade, P. (1971). Evaluación de los aprendizajes. Buenos Aires (Argentina), Kapelusz.

Laitenberger, O. y J. DeBaud (1997). "Prespective-base reading of code documents at Robert Boshsc GmbH." Information and Software Technology **39**(11): 781-791.

Landeta, J. (1999). El método Delphi. Una técnica de previsión para la incertidumbre. Barcelona, Ariel.

Landeta, J. (2006). "Current Validity of the Delphi method in social sciences." Technological Forecasting and Social Change **73**: 467-482.

Landis, J. R. y G. G. Koch (1977). "The measurement of observer agreement for categorical data." Biometrics **33**: 159-174.

Lang, T. (1995). "An Overview of Four Futures Methodologies (Delphi, Environmental Scanning, Issues Management and Emerging Issues Analysis)." The Manoa Journal of Fried and Half-Fried Ideas (about the future) **7**: 28.

Lazarfeld, P. (1950). The logical and mathematical foundationes of latent structure analysis. Princenton, Princenton University Press.

Linstone, H. A. y M. Turoff (1975). The Delphi method: Techniques and Aplications. london, Addison Wesley Publishing.

López-Cuadrado, J. (2003). The CAT is out of the bank. En A. M. Vilas, J. A. M. González y J. M. González (eds.) Proc. of Advances in technology-based education: towards a knowledge-based society, Badajoz (España), Junta de Extremadura (CECT), **3**: 1832-1836.

López-Cuadrado, J. (2006). Administración de subtests de anclaje para calibrar un banco de ítems. San Sebastián (España), University of the Basque Country (UPV/EHU/LSI/TR 8-2006): 96.

López-Cuadrado, J. (2008). Evaluación mediante test adaptativos informatizados en el contexto de un sistema adaptativo para el aprendizaje de la lengua vasca. Lenguajes y Sistemas Informáticos. San Sebastián, Univ. País Vasco/ Euskal Herriko Unibertsitates**:** 401.

López-Cuadrado, J. y A. J. Armendariz (2006). Obtención de estimaciones de los parámetros durante la calibración de un banco de ítems. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 13-2006): 271.

López-Cuadrado, J., A. J. Armendariz y T. A. Pérez-Fernández, Eds. (2005). A supporting tool for the adaptive assessment of an e-learning system. Recent Research Developments in Learning Technologies. Cáceres (Spain), Formatex Research Center-Badajoz.

López-Cuadrado, J., A. J. Armendariz, T. A. Pérez y R. Arruabarrena (2008). Helping Tools For Item Bank Calibration And Development Of Computrized Adaptive Tests. En L. G. Chova, D. M. Belenguer y I. C. Torres (eds.) Proc. of International Technology, Education and Development Conference (INTED'08), Valencia (España), International Association of Technology, Education And Development (IATED): 1-9.

López-Cuadrado, J., A. J. Armendariz, T. A. Pérez, R. Arruabarrena y J. A. Vadillo (2009). Chapter "Computerized adaptive testing, the item bank calibration and a tool for easing the process." International Technology, Education and Development Conference: 1-22.

López-Cuadrado, J. y R. Arruabarrena (2005). Diseño de anclaje de un banco de ítems. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 12-2005): 109.

López-Cuadrado, J., R. Arruabarrena y A. J. Armendariz (2005). La Salle Andoaingo Ikastolako ikasleen euskara frogen emaitzen txostena. San Sebastián, University of the Basque Country (UPV/EHU/LSI/TR 25-2005): 40.

López-Cuadrado, J., T. A. Pérez, J. A. Vadillo y R. Arruabarrena (2002). Integrating Adaptive Testing in an Educational System. En E. Kähkönen y E. Sutinen (eds.) Proc. of Educational Technology in Cultural Context: ETCC2002. First International Conference on, Joensuu, Finland, (Joensuun Yliopisto, Intenational Proceeding Series), University of Joensuu: 133-139.

López-Cuadrado, J., T. A. Pérez, J. A. Vadillo y J. Gutiérrez (in press). "Calibration o fan item bank for the assessment of Basque language knowledge." Computers & education: doi:10.1016/j.compedu.2010.04.015.

López-Cuadrado, J., M. Villamañe, J. Gutiérrez, T. A. Pérez, R. Arruabarrena y J. A. Vadillo (2001). CiberBiblio: una biblioteca virtual multimedia. En P. d. l. F. R. y. A. P. Alarcón (eds.) Proc. of II Jornadas Españolas de Bibliotecas Digitales: JIBIDI'01, Almagro, Ciudad Real (España): 12.

López Pina, J. A. (1995). Teoría de la respuesta al ítem: fundamentos. Murcia (España), Barcelona: Promociones y Publicaciones Universitarias (PPU); Mucia DM.

López Pina, J. A. y M. D. Hidalgo Montesinos (1996). Chapter "Bondad de ajuste y teoría de respuesta de los ítems." Psicometría. J. Muñiz. Madrid (España), Editorial Universitas, S.A.: 643-704.

López_de_Ullibarri, I. y S. Pita (1999). "Medidas de concordancia: el índice Kappa." CAD ATEN PRIMARIA. Metodología de la Investigación **6**: 169-171.

Lord, F. M. (1952). "A theory of test scores." Psychometric Monograph **7**.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, N.J., L. Erlbaum Associates.

Lord, F. M. y M. Novick (1968). Statistical Theories of Mental Tests Scores. New York, Addison Wesley.

Loyd, B. H. y H. D. Hoover (1980). "Vertical equating using the Rasch model." J. Educational Measurement **17**: 179-193.

Lumsden, J. (1976). "Test theory." Annual Review of PsychologyAnnual Review of Psychology **27**: 251-280.

Manly, B. F. J. (1994). Multivariate Statistics Methods - A Primer. 2nd edition. London (Reino Unido), Chapman & Hall.

Marco, G. L. (1977). "Item characteristic curve solutions to three intractable testing problems." J. Educational Measurement **14**: 139-160.

Mark, M. A. y J. E. Greer (1993). "Evaluation Methodologies for Intelligent Tutoring Systems." International Journal of Artificial Intelligence in Education. Special Issue on Evaluation **4**(2/3): 129-153.

Martínez-Cervantes, R. J. y R. Moreno-Rodríguez (2002). Construcción de un banco de ítems informatizado para la evaluación de conocimientos sobre una materia universitaria. En J. M. d. Mesa, R. Castañeda y L. M. Villar (eds.) Proc. of Proc. III Jornadas Andaluzas de Calidad en la Enseñanza Universitaria, Universidad de Sevilla, **vol. II**: 329-344.

McGraw, K. L. y K. Harbison-Briggs (1989). "Knowledge acquisition: Principles and guidelaines."

MEC (2007). PIRLS 2006. Informe español. En Catálogo oficial de publicaciones oficiales (NIPO: 651-07-388-0; ISBN: 978-84-369-4528-7): pp.107.

Miller, J., M. Wood y M. Roper (1998). Further experiences with Scenarios and Checklists.

Mislevy, R. J. (1986). "Bayes modal estimation in item response models." J. Psychometrika **51**: 177-195.

Molenaar, I. W. (1995). Chapter "Estimation of item parameters." Rasch models: foundations, recent developments, and applications. G. H. Fischer y I. W. Molenaar. New York (USA), Springer-Verlag: 39-51.

Mulaik, S. A., N. S. Raju y R. A. Harshman (1997). Chapter "There is a time and a place for significance testing." What if there were no significance tests? S. A. M. Lisa L. Harlow, and James H. Steiger. Mahwah, NJ (EE.UU.), Lawrence Erlbaum: 65-115.

Mullis, I. V. S., M. O. Martin y A. M. Kennedy (2007a). PIRLS 2006. Technical Report. Boston College, MA 02467 (United States), TIMSS & PIRLS Intenational Study Center - IEA.

Mullis, I. V. S., M. O. Martin, A. M. Kennedy y P. Foy (2007b). PIRLS 2006. International Report. Boston College, MA 02467 (United States), TIMSS & PIRLS Intenational Study Center - IEA.

Muñiz, J. (1992). Teoría clásica de los Tests. Madrid, Ediciciones Pirámide.

Muñiz, J. (1996). Psicometría. Madrid (España), Editorial Universitas, S.A.

Muñiz, J. (1997). Introducción a la teoría de respuesta a los ítems. Madrid (España), Ediciones Pirámide.

Muraki, E., C. M. Hombo y Y.-W. Lee (2000). "Equating and linking of performance assessments." J. Applied Psychological Measurement **24**(4): 325-337.

Murray, T. (1993). "Formative Qualitative Evaluation for "Exploratory" ITS Research." International Journal of Artificial Intelligence in Education. Special Issue on Evaluation **4**(2/3): 179-207.

Navas, M. J. (1996). Chapter "Equiparación de puntuaciones." Psicometría. J. Muñiz. Madrid (España), Editorial Universitas, S.A.: 293-369.

Nielsen, J. (1993). Usability Engineering, AP Professional (Academic Press).

Nielsen, J. y R. L. Mack (1994). Usability Inspection Methods. New York, John Wiley &Sons, Inc.

Nitko, A. J. y T. C. Hsu (1984). "A comprehensive microcomputer system for classroom testing." Journal of Educational Measurement **21**: 377-390.

Norman, D. A. (1983). Chapter "Some observations on mental models." Mental models. D. Gentner y A. L. Stevens. Hillsdale, New Jersey (USA), Lawrence Erlbaum Associates: 7-14.

O'Brien, M. L. y J. O. Hampilos (1988). "The feasibility of creating an item bank from a teacher-made test using the Rash model." Educational and Psychological Measurement(48): 201-212.

O'Hanlon, N. (1999). "Web-Based Tutorials: Does a Course Use Differ From General Use?" International Journal of Interactive Learning Research **10**(2): 217-228.

OCDE (2003). PISA 2003 Assessment Framework - Mathematics, Reading, Science and Problem Solving Knowledge and Skills: pp.194.

OCDE (2005). PISA 2003 Technical Report: (Complete Edition), Organization for Economic Cooperation & Development.

Ogasawara, H. (2001). "Standard errors of item response theory equating/ linking by response function methods." J. Applied Psychological Measurement **25**(1): 53-67.

Olea, J., F. J. Abad y V. Ponsoda (2002). "Elaboración de un banco de ítems, predicción de la dificultad y diseño de anclaje." Metodología de las Ciencias del Comportamiento **Volumen Especial**: 427-430.

Olea, J. y P. Hontangas (1999). Chapter "Tests informatizados de primera generación." Tests informatizados: fundamentos y aplicaciones. J. Olea, V. Ponsoda y G. Prieto. Madrid (España), Ediciones Pirámide: 111-125.

Olea, J. y V. Ponsoda (2003). Tests adaptativos informatizados. Madrid (España), Ediciones UNED.

Olea, J., V. Ponsoda, J. Revuelta y J. Belchi (1996). "Propiedades psicométricas de un test adaptativo informatizado de vocabulario inglés." Estudios de Psicología **55**: 61-73.

Or-Bach, R. y E. Bar-On (1993). "TALK-ing About Evaluation." International Journal of Artificial Intelligence in Education **4**(2/3): 227-243.

Otero, M. C. (2003). Evaluación empírica de la comprensión del modelado dinámico en los lenguajes UML y OML de aplicaciones software. Lenguajes y Sistemas Informáticos, UPV/EHU**:** 310.

Ozen, D. J. y S. P. Reise (1994). Personality assessment. Palo Alto.

Parlett, M. R. y D. Haminlton (1987). "Evaluating education: issues and methods." **1**(4): 57-73.

Parry, J. D. y A. M. Hofmeister (1986). "Tthe development and validation of an expert system for special education." Computational Intelligence **2**: 65-67.

Patridge, D. (1986). Artificial intelligence: Applications in the future of software engineering, NewYork: EllisHorwood.

Pérez, C. (1999). Técnicas de muestreo estadístcio. Teoría, práctica y aplicaciones informáticas. Madrid (Spain), ra-ma.

Pérez, T. A. (2000). Un hiperentorno adaptativo para el aprendizaje instructivo / constructivo. Lenguajes y Sistemas Informáticos. San Sebastián, Uinv. País Vasco / Euskal Herriko Univertsitatea**:** 286.

Pérez, T. A., K. Gabiola, J. Gutiérrez, R. López, A. González y J. A. Carro (1999). Hezinet: Interactive (Adaptive) Education Through Activities. En (eds.) Proc. of Educational Multimedia, Hypermedia and Telecommunications, ED-MEDIA, Seattle, USA, AACE, **Addendum**:

Pérez, T. A., J. Gutiérrez, R. López, A. González y J. A. Vadillo (2001). "Hipermedia, adaptación, constructivismo e instructivismo." Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial(12): 29-38.

Pérez, T. A., J. Gutiérrez y P. Lopistéguy (1995a). The role of exercise in a usre-adapted hypermedia. En (eds.) Proc. of 3rd Computer Aided Engineering Education (CAEE'95), Bratislava (Slovakia):

Pérez, T. A., J. Gutiérrez, P. Lopistéguy y I. Usandizaga (1995b). The Role of Exercices in a User Adaptive Hypermedia. En (eds.) Proc. of International Conference on Computer Aided Engineering Education, CAEE, Bratislava, Eslovaquia: 57-62.

Pérez, T. A., R. López, J. Gutiérrez y A. González (2000). Chapter "Learning Basque in a Distance-adaptive way." Computers and Education in the 21st Century. M. Ortega y J. Bravo. Dordrecht, The Netherlands, Kluwer Academic Publishers: 251-262.

Plaza, I., J. J. Marcuello y F. J. Arcega (2007). Evaluación de la calidad del software educativo: revisión normativa. En I. F. D. Castro (eds.) Proc. of Simposio Nacional de Tecnologías de la Información y las Comunicaciones en la Educacion: SIntice-2007, incluido en el II Congreso Español De Informática: CEDI'07(SINTICE-CEDI'07), Zaragoza, Nuevos retos científicos y tecnológicos en Ingeniería Informática, Thomsom: 197-204.

Pollit, A. B. (1985). Chapter "Item banking and school measurement." New Directions in Educational Psychology: Learning and teaching. W. Entwistle. East Sussex (England), The Falmer Press.

Porter, A. A., L. G. Votta y V. R. Basili (1995). "Comparing detection methods for software requirements inspections: a replicated experiment." IEEE Transactions on Software Engineering **21**(6): 563-575.

Reid, A. y M. Arends (1998). Evaluation of Computer-Assisted Learning Program Question Styles and Integration into a General Pathology Course, LTDI Evaluation Studies (Learning Technology Dissemination Initiative).

Renom, J. (1993). Tests adaptativos computerizados: fundamentos y aplicaciones. Barcelona (España), Promociones y Publicaciones Universitarias.

Renom, J. y E. Doval (1999). Chapter "Tests adaptativos informatizados: estructura y desarrollo." Tests informatizados: fundamentos y aplicaciones. J. Olea, V. Ponsoda y G. Prieto. Madrid (España), Ediciones Pirámide: 127-162.

Rolón, E., F. García, F. Ruiz y M. Piattini (2007). Familia de experimentos para validar medidas para Modelos de Procesos de Negocio con BPMN. En F. Ruiz y F. O. García (eds.) Proc. of Proceedings of the I Taller sobre Procesos de Negocio e Ingeniería del Software (electronic proceedings). Actas de los Talleres de las XII Jornadas de Ingeniería del Software y Bases de Datos (TJISBD). Zaragoza (España), **vol.1;no.2**: 41-48.

Santisteban, C. y J. M. Alvarado (2001). Modelos psicométricos. Madrid (España), Ediciones UNED.

Sanz-Lumbier, S., J. Gutiérrez, T. A. Pérez, S. Sanz-Santamaría, J. A. Vadillo y M. Villamañe (2002). Hezinet. The Hypermedia System That Makes The Basque Language Easy To Learn. En Gustavo A. Santana Torrellas y V. Uskov (eds.) Proc. of IASTED: 5th International Conference on Computers and Advanced Technology in Education, Cancún (Mexico), ACTA Press: 344-349.

Scriven, M. (1991). Evaluation Thesaurus. 4th edition, Sage Publications.

Schmidt, F. L. (1996). "Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers." J. Psychological Methods **1**: 115-129.

Schofield, J. W., D. Evans-Rhodes y B. R. Huber (1990). "Artificial intelligence in the classroom: The impact of a computer-based tutor on teachers and students." Social Science Computer Review **8**(1): 24-41.

Schwarz, E., P. Brusilovsky y G. Weber (1996). World-wide intelligent textbooks. En (eds.) Proc. of 1st World Conference on Educational Telecommunications (ED-TELECOM'96), Boston, MA (EE.UU), Association for the Advancement of Computing in Education (AACE): 302-307.

Shneiderman, B. (1998). Designing the User Interface, 3rd edition. Reading-Massachusetts, Addison Wesley Longmang, Inc.

Shute, V. J. y W. Regian (1993). "Principles for Evaluating Intelligent Tutoring Systems." International Journal of Artificial Intelligence in Education. Special Issue on Evaluation **4**(2/3): 245-271.

Sjøberg, S. (2004). Science Education: The voice of the learners. En (eds.) Proc. of Proc. Conference on Increasing Human Resources for Science and Technology in Europe, Bruselas (UE):

Stackman, H. (1974). Delphi assessment: expert opinio, forecasting, and group process. Santa Monica. CA (EE.UU), Rand Corporation:

Stage, C. (2003). "Teoría clásica de medición o teoría de respuesta al ítem. La experiencia sueca." eJournal Estudios Públicos **nº 90**: 185-217.

Swaminathan, H., R. K. Hambleton, S. G. Sireci, D. Xing y S. M. Rizavi (2003). "Small sample estimation in dichotomous item response models: effects of priors based on judgmental information on the accuracy of item parameter estimates." J. Applied Psychological Measurement **27**(1): 27-51.

Tapias, J. A. (2008). 1.112 preguntas básicas para oposiciones a bombero, Ed. Cep.

Taylor, J., M. Woodman, T. Summer y C. T. Blake (2000). "Peering Through a Glass Darkly: Integrative evaluation of an on-line course." Educational Techonology & Society **3**(4).

Tessmer, M. (1993). Planning and Conducting: Formative Evaluations. London, Kogan Page Limited.

Teusch, P., T. Chanier, Y. Chevalier, D. Perrin, F. Mangenot, J. P. Narcy y J. De Saint Ferjeux (1996). Chapter "Environnements interactifs pour l'apprentissage en langue étrangère." Hypermedias et apprentissages. E. Bruillard, M. Baldner y L. Baron, INRP-EPI: 247-256.

Tognolini, J. (1982). Pupil achievement in stage 6 mathematics. (Discussion paper N. 15), Perth: Education Department of Western Australia:

Traub, R. E. y Y. R. Lam (1985). "Latent structure and item sampling models for testing." Annual Reviews Of Psychology **36**: 19-48.

Tucker, L. R. (1946). "Maximum validity of a test with equivalent items." Psychometrika **11**: 1-13.

Tukey, J. W. (1986). Chapter "The future of data analysis." The Collected Works of John W Tukey. Vol. III: Philosophy and Principles of Data Analysis 1949-1964, Chapman & Hall/CRC: 391-484.

Tuya, J., I. Ramos y J. Dolado (2007). Técnicas cuantitativas para la gestión en la ingeniería del software. Oleiros-A Coruña (España), NETBIBLO.

Vale, C. D., V. A. Maurelli, K. A. Gialluca, D. J. Weiss y M. J. Ree (1981). Methods for linking item parameters. TX (USA), Air Force Human Resources Laboratory - Brooks Air Force Base:

van Solinger, R. y E. Berghoout (1999). The Goal/Question/Metric. A practical guide for quality improvement of Software Development, McGraw-Hill Education.

van Thiel, C. C. y M. A. Zwarts (1986). "Development of a Testing Service System." Journal of Applied Psychological Measurement **10**(4): 391-403.

Villamañe, M., J. Gutiérrez, R. Arruabarrena, T. A. Pérez, S. Sanz-Lumbier, S. Sanz-Santamaría y J. López-Cuadrado (2001). Use and Evaluation of HEZINET: a system for Basque language learning. En (eds.) Proc. of ICCE, Seoul, South Korea, AACE: 93-101.

Virvou, M. y B. du Boulay (1999). "Human palusible reasoning for intelligent help." User Modeling and User-Adapted Interaction **9**(4): 323--377.

Wackerly, D. (2002). Estadistica Matematica Con Aplicaciones - 6ª ed., I.T.P. Latin America.

Wainer, H., N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg y D. M. Thissen (2000). Computerized Adaptive Testing: A Primer (2nd edition). Hillsdale, New Jersey, Lawrence Erlabum Associates.

Wainer, H., N. J. Dorans, B. F. Green, R. J. Mislevy, L. Steinberg y D. M. Thissen (1990). Chapter "Future challenges." Computerized adaptive testing: a primer. H. Wainer. Hillsdale, New Jersey (USA), Lawrence Erlbaum Associates: 65-102.

Wainer, H. y R. J. Mislevy (1990). Chapter "Item response theory, item calibration and proficiency estimation." Computerized adaptive testing: a primer. H. Wainer. Hillsdale, New Jersey (USA), Lawrence Erlbaum Associates: 65-102.

Wainer, H. y R. J. Mislevy (2000). Chapter "Item response theory, item calibration, and proficiency estimation." Computerized adaptive testing: a primer (second edition). H. Wainer. Mahwah, New Jersey (USA), Lawrence Erlbaum Associates: 61-99.

Weber, G., H.-C. Kuhl y S. Weibelzahl (2001). Chapter "Developing adaptive internet based course with the authoring system NetCoach." LNCS. Revised Papers from the nternational Workshops OHS-7, SC-3, and AH-3 on Hypermedia: Openness, Structural Awareness, and Adaptivity. S. Reich, M. Tzagarakis y P. De Bra. London (UK), Springer-Verlag: 226-238.

Weibelzahl, S. (2002). Evaluation of Adaptive Systems. (Ph dissertation). Faculty I. Freiburg (Alemania), University of Trier**:** 169.

Weiss, D. J. y M. E. Yoes (1991). Chapter "Item response theory." Advances in Educational and Psychological Testing: Theory and Applications. R. K. Hambleton y J. N. Zaal. Norwell (Netherlands), Kluwe Academic: 69-95.

Welch, M. y K. Brownell (2000). "The Development and Evaluation of a Multimedia Course on Educational Collaboration." International Journal of Educational Multimedia and Hypermedia **9**(3): 169-194.

White, S. A. (2004). Introduction to BPMN. BPTrends (june, 2004)**:** pp. 21 http://www.bptrends.com/publicationfiles/07-04%20WP%20Intro%20to%20BPMN%20-%20White.pdf.

Wise, S. L. y G. G. Kingsbury (2000). "Practical issues in developing and maintaining a computerized adaptive testing program." J. Psicológica **21**: 135-155.

Wohlin, C., P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell y A. Wesslén (2000). Experimentation in Software Engineering: An Introduction. Massachusetts (EE.UU), Kluwer Academic Publishers.

Wonnacott, R. J. y T. H. Wonnacott (1991). Estadística Básica Práctica. Su utilidad y múltiples aplicaciones. México, LIMUSA.

Worthen, B. R., J. R. Sanders y J. L. Fitzpatrick (1997). Program Evaluation. Alternative Approaches and Practical Guidelines,2nd ed. New York (EE.UU.), Addison Wesley Longman.

Wright, B. D. y S. R. Bell (1984). "Items banks: what, why and how." Journal of Educational Measurement **21**(4): 331-346.

Zuse, H. (1998). A framework of Software Measurement. Berlin (Alemania), Walter de Gruyter.