Departamento de
Lenguajes y Sistemas Informáticos

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Calibración de ítems mediante juicio de expertos utilizando técnicas de ingeniería dirigida por modelos, workflows y sistemas de gestión de aprendizaje

MEMORIA

Que para optar al grado de Doctor en Informática presenta

Mª Concepción Presedo García

Bilbao, 2 de Febrero de 2017

AGRADECIMIENTOS

Estimado lector,

Antes de dar comienzo a esta memoria quisiera expresar mi más sincera gratitud a todos aquéllos que, en mayor o menor medida, han facilitado y hecho posible la culminación de este trabajo de investigación.

Gracias a todos.

En Bilbao, a 2 de Febero de 2017,

Conchi.

Índices y Glosario

Índice de contenidos

ÍNDICES Y GLOSARIO **I**

ÍNDICE DE CONTENIDOS **I**

ÍNDICE DE ECUACIONES **V**

ÍNDICE DE FIGURAS **VII**

ÍNDICE DE TABLAS **IX**

ABREVIATURAS Y ACRÓNIMOS **XI**

PARTE PRIMERA: INTRODUCCIÓN **1**

I OBJETIVOS Y CONTEXTO **5**

I 1 CONTEXTO DEL TRABAJO REALIZADO 6

I 2 REQUERIMIENTOS Y OBJETIVOS 9

I 3 ORGANIZACIÓN DE LA MEMORIA Y GUÍA DE LECTURA 11

II ANTECEDENTES: CALIBRACIÓN DE LOS ÍTEMS DE HEZINET **13**

II 1 DISEÑO DEL EXPERIMENTO 13

II 2 ADMINISTRACIÓN DE LOS ÍTEMS 15

II 3 ANÁLISIS DE DATOS Y CALIBRACIÓN 15

PARTE SEGUNDA: FUNDAMENTOS **19**

III PSICOMETRÍA BÁSICA **23**

III 1 EL TEST COMO MEDIDOR DE RASGO O HABILIDAD 23

III 2 TEORÍAS PSICOMÉTRICAS 24

III 2.1 La Teoría Clásica de los Test (TCT) 24

III 2.2 La Teoría de Respuesta al Ítem (TRI) 26

III 2.3 Diferencias entre TCT y TRI 30

IV ESTANDARIZACIÓN DE RECURSOS EDUCATIVOS **33**

IV 1 INICIATIVAS DE ESTÁNDARES E-LEARNING 33

IV 2 MARCO DE TRABAJO IMS 35

IV 3 ESTÁNDARES IMS PARA EL EMPAQUETADO DE CONTENIDOS 37

IV 4 ESTÁNDAR IMS PARA EL MODELADO DE ÍTEMS Y TEST 38

V BANCOS DE ÍTEMS CALIBRADOS **43**

V 1 ÁMBITOS DE APLICACIÓN 43

V 2 UTILIDAD EN LOS SISTEMAS DE APRENDIZAJE 46
V 3 CALIBRACIÓN DE ÍTEMS BASADA EN LA TRI 47

VI PROCESO DE CALIBRACIÓN MEDIANTE EL JUICIO DE EXPERTOS 51
VI 1 FUNDAMENTOS SOBRE LA CALIBRACIÓN DE ÍTEMS CON EXPERTOS 51
VI 2 PROCESO DE CALIBRACIÓN DE ÍTEMS VÍA EXPERTOS 55
VI 3 DISEÑO DEL PROCESO DE CALIBRACIÓN 58
VI 3.1 Especificación de ítems y experimento 59
VI 3.2 Especificación de análisis, estudios y cálculos para calibración 61

PARTE TERCERA: ESTADO DEL ARTE 65

VII SOFTWARE PARA ADMINISTRACIÓN 69
VII 1 REVISIÓN DE PLATAFORMAS LCMS CON ESTÁNDARES E-LEARNING 72
VII 2 REVISIÓN DETALLADA DE LA PLATAFORMA MOODLE 76

VIII SOFTWARE PARA ANÁLISIS DE DATOS Y CALIBRACIÓN 79
VIII 1 REVISIÓN DEL SOFTWARE 80

PARTE CUARTA: LA HERRAMIENTA DE AYUDA CALLIE-EXPERT 89

IX EL SISTEMA CALLIE-EXPERT 93
IX 1 METAMODELOS PARA LA ESPECIFICACIÓN DEL PROCESO 93
IX 1.1 Metamodelo de calibración 95
IX 1.2 Metamodelo de cuestionarios 97
IX 1.3 Metamodelo de análisis y cálculos 98
IX 2 METAMODELOS PARA EL INTERCAMBIO DE DATOS Y RESULTADOS 99
IX 2.1 Metamodelo de aportaciones 99
IX 2.2 Metamodelo de resultados 101
IX 3 ARQUITECTURA DE CALLIE-EXPERT 103
IX 3.1 CALLIE-EXPERT: ESKARI 104
IX 3.2 CALLIE-EXPERT: ADMINQ Factory 106
IX 3.3 CALLIE-MOODLE: Curso y cuentas 108
IX 3.4 CALLIE-EXPERT y WWF: WF Factory y Workflow CA 109
IX 3.5 CALLIE-EXPERT: PRO-C 110
IX 4 EL MÓDULO EXT 111

X INTERFAZ DE CALLIE-EXPERT 113
X 1 INTERFAZ DE ESKARI: CALLIE-ESKARI 114
X 2 INTERFAZ DE PRO-C: CALLIE-PRO 124

XI EVALUACIÓN DE CALLIE-EXPERT 133
XI 1 PRUEBAS DE COMPONENTES CON USUARIOS 134
XI 1.1 Pruebas de introducción de ítems con ESKARI 135
XI 1.2 Pruebas de diseño del experimento con ESKARI 135
XI 1.3 Pruebas con CALLIE-MOODLE 137
XI 1.4 Pruebas con Excel para PRO-C 138

XI 1.5 Pruebas con PRO-C 139
XI 1.6 Resultados globales para los componentes 140
XI 2 RESULTADOS OBTENIDOS POR CALLIE-EXPERT 142
XI 2.1 Réplica de la calibración de los ítems de Hezinet 142
XI 2.2 Réplica de la calibración de los ítems de IRALE 146
XI 3 PRUEBAS GLOBALES PARA LA APLICACIÓN WEB 149
XI 4 PRUEBAS DE INTEGRACIÓN CON CALLIE-TRI 150

XII CONCLUSIONES 153
XII 1 PRINCIPALES APORTACIONES 153
XII 2 PRINCIPALES LÍNEAS FUTURAS DE TRABAJO 156
XII 3 PUBLICACIONES 159

PARTE QUINTA: ANEXOS Y BIBLIOGRAFÍA 161

A1 EJEMPLO DE MODELO DE CALIBRACIÓN 165
A2 EXPERIMENTO 1: CALIBRACIÓN DE LOS ÍTEMS DEL TEMA IRALE 167
A3 EXPERIMENTO 2: RÉPLICAS Y EVALUACIÓN CON ALUMNOS 185
A4 INFORME DE RESULTADOS RÉPLICA HEZINET 193
A5 INFORME DE RESULTADOS RÉPLICA IRALE 209
REFERENCIAS BIBLIOGRÁFICAS 221

Índice de ecuaciones

Ecuación 1 – Relación lineal entre puntuación real y observada. *24*
Ecuación 2 – Modelo logístico de un parámetro. *29*
Ecuación 3 – Modelo logístico de tres parámetros (Birnbaum, 1968). *29*

Índice de figuras

Figura 1 – Detalle de un ítem del apartado del cuestionario Ítems a valorar. *14*
Figura 2 – Algoritmo de aplicación de los filtros en el experimento Hezinet. *16*
Figura 3 – Ejemplo de CCI para un modelo logístico de tres parámetros. *27*
Figura 4 – El marco de contenidos IMS (IMS-CP, 2001). *35*
Figura 5 – Modelo para el manifiesto de un paquete IMS CC v1.0.0. (IMS-CC, 2017) *37*
Figura 6 – Metamodelo ASI para el estándar IMS QTI (Smythe et al., 2002) *39*
Figura 7 – Ejemplo de ítem en formato IMS QTI *40*
Figura 8 – Ítem de selección múltiple EUSK1 con una sola respuesta correcta. *41*
Figura 9 – Metamodelo de Calibraciones de CALLIE-EXPERT. *94*
Figura 10 - Diagrama de estados del proceso de calibración en CALLIE-EXPERT. *94*
Figura 11 – Metamodelo de calibración MMCA. *96*
Figura 12 – Metamodelo de cuestionarios MMCU. *97*
Figura 13 – Metamodelo de análisis y cálculos MMANCA. *98*
Figura 14 – Metamodelo de aportaciones MMAP. *100*
Figura 15 – Metamodelo de resultados MMRE. *101*
Figura 16 – Arquitectura del sistema de calibración. *103*
Figura 17 – Página de Moodle correspondiente a un curso de administración. *107*
Figura 18 – Presentación de un cuestionario para calibración vía expertos. *107*
Figura 19 – Página de Moodle correspondiente a los participantes del curso. *108*
Figura 20 – Arquitectura del workflow CA. *110*
Figura 21 – Mapa con los pasos del asistente para CALLIE-ESKARI. *115*
Figura 22 – CALLIE-ESKARI: Página de bienvenida. *115*
Figura 23 – CALLIE-ESKARI: Página de Menú de pasos. *116*
Figura 24 – CALLIE-ESKARI: Página de especificación del tema a calibrar. *116*
Figura 25 – CALLIE-ESKARI: Página de inclusión de ítems. *117*
Figura 26 – CALLIE-ESKARI: Página de introducción de ítems de selección múltiple en IMS QTI. *117*
Figura 27 – CALLIE-ESKARI: Página de especificación del tipo de calibración. *118*
Figura 28 – CALLIE-ESKARI: Página de configuración inicial para la administración. *118*
Figura 29 – CALLIE-ESKARI: Página de especificación de datos sobre los expertos. *119*
Figura 30 – CALLIE-ESKARI: Página de especificación de diseño homogéneo. *119*
Figura 31 – CALLIE-ESKARI: Página de especificación de diseño no homogéneo. *120*
Figura 32 – CALLIE-ESKARI: Página de definición de bloques. *120*
Figura 33 – CALLIE-ESKARI: Página de reparto de ítems. *121*
Figura 34 – CALLIE-ESKARI: Página de otras decisiones de administración y filtrado. *122*
Figura 35 – CALLIE-ESKARI: Menú de pasos con los cuatro pasos dados. *122*
Figura 36 – CALLIE-ESKARI: Página ver resumen. *123*
Figura 37 – CALLIE-ESKARI: Página de petición aceptada. *123*
Figura 38 – Mapa con los pasos del asistente para CALLIE-PRO. *124*
Figura 39 – CALLIE-PRO: Página de bienvenida. *125*
Figura 40 – CALLIE-PRO: Página procesar mis peticiones de calibración. *125*
Figura 41 – CALLIE-PRO: Página ver resumen progreso. *126*
Figura 42 – CALLIE-PRO: Web de CALLIE-MOODLE. *127*
Figura 43 – CALLIE-PRO: Distintas opciones posibles para una petición no concluida. *127*
Figura 44 – CALLIE-PRO: Página generar Excel. *128*
Figura 45 – CALLIE-PRO: Página ver resultado de la calibración. *128*
Figura 46 – CALLIE-PRO: Informe de resultados generado en la descarga MS Excel. *129*
Figura 47 – CALLIE-PRO: Página cambio de filtros y simulación. *130*
Figura 48 – CALLIE-PRO: Página opciones de finalización. *130*
Figura 49 – CALLIE-PRO: Página generar banco calibrado. *131*
Figura 50 – Resultados globales por funcionalidad. *141*

Figura 51 – Resultados globales por componente. 141
Figura 52 – Resumen de los resultados de Hezinet obtenidos por CALLIE-EXPERT. 145
Figura 53 – Resumen de los resultados de IRALE obtenidos por CALLIE-EXPERT. 149
Figura 54 – Resultados globales para la aplicación Web. 150
Figura 55 – XML de petición de una calibración tras la especificación del experimento. 165
Figura 56 – Formato de un ítem de IRALE para evaluar la capacidad "comprensión lectora". 167
Figura 57 – Escala CEFR de 11 niveles utilizada para la calibración mediante expertos. 168
Figura 58 – Datos a rellenar por ítem en la prueba 3 con CALLIE-MOODLE. 177

Índice de tablas

Tabla 1 – Diferencias entre la TCT y la TRI (Muñiz, 2010). *30*
Tabla 2 – Pasos en la calibración de ítems basada en la TRI. *48*
Tabla 3 – Agentes involucrados en la calibración de ítems. *55*
Tabla 4 – Tareas principales del proceso integral de calibración. *56*
Tabla 5 – Tareas en el diseño del experimento a realizar. *56*
Tabla 6 – Tarea en la ejecución de la fase de administración vía expertos. *57*
Tabla 7 – Tareas en la ejecución de la fase de análisis y calibración vía expertos. *58*
Tabla 8 – Síntesis de compañías/proyectos que soportan estándares e-learning. *76*
Tabla 9 – Síntesis de programas de apoyo para el análisis y calibración de ítems. *87*
Tabla 10 – Resumen de las pruebas realizadas para la introducción de ítems. *135*
Tabla 11 – Características de cada una de las calibraciones a diseñar. *136*
Tabla 12 – Resumen de resultados de las pruebas para el diseño del experimento. *136*
Tabla 13 – Resumen de la prueba CALLIE-MOODLE parte responsables de la calibración. *137*
Tabla 14 – Resumen de la prueba CALLIE-MOODLE parte expertos participantes. *138*
Tabla 15 – Resumen de la prueba con Excel. *138*
Tabla 16 – Resumen de la prueba PRO-C sobre el progreso del proceso. *139*
Tabla 17 – Revisión de la prueba PRO-C sobre la discusión de resultados. *140*
Tabla 18 – Puntuaciones medias por funcionalidad y grupo para CALLIE-EXPERT. *141*
Tabla 19 – Características de diseño para la réplica de Hezinet. *143*
Tabla 20 – Evolución de la muestra durante el análisis de CALLIE-EXPERT para Hezinet. *143*
Tabla 21 – Ítems extra retirados por el filtro C.it-2 en CALLIE-EXPERT. *144*
Tabla 22 – Características iniciales de diseño para la réplica de IRALE. *147*
Tabla 23 – Evolución de la muestra durante el análisis inicial de CALLIE para IRALE. *147*
Tabla 24 – Resultados por grupos para la aplicación Web. *150*
Tabla 25 – Ítems del banco de IRALE clasificados por los responsables. *170*
Tabla 26 – Resumen de contribuciones por experto. *171*
Tabla 27 – Informe de los datos recogidos por los responsables de IRALE. *174*
Tabla 28 – Distribución final de los ítems del banco según los 11 niveles del CEFR. *174*
Tabla 29 – Bloques e ítems asignados a cada bloque. *176*
Tabla 30 – Cuestionarios, bloques de cada uno y asignación a expertos. *176*
Tabla 31 – Cuestionario a cumplimentar por cada experto del experimento 1. ..... *180*
Tabla 32 – Revisión de las tareas de los responsables en el experimento 1. ..... *183*
Tabla 33 – Revisión de las tareas de los participantes expertos en el experimento 1. ..... *183*
Tabla 34 – Correspondencias entre los 11 niveles del CEFR y los niveles para CALLIE-EXPERT. *189*
Tabla 35 – Revisión de tiempos y tareas de los participantes en el experimento 2. ..... *192*
Tabla 36 – Detalle de los resultados para los expertos en el caso Hezinet. *196*
Tabla 37 – Frecuencias utilizadas en el cálculo de la dificultad en el caso Hezinet. *200*
Tabla 38 – Cálculo detallado de la dificultad en los 20 ítems ambiguos en el caso Hezinet. *201*
Tabla 39 – Detalle de los resultados de calibración para los 252 ítems de Hezinet. *208*
Tabla 40 – Detalle de los resultados para los expertos en IRALE. *209*
Tabla 41 – Frecuencias utilizadas en el cálculo de la dificultad en el caso 1 de IRALE. *213*
Tabla 42 – Cálculo detallado de la dificultad en el ítem ambiguo en el caso 1 de IRALE. *213*
Tabla 43 – Detalle de los resultados para los 132 ítems en el caso 1 de IRALE. *216*
Tabla 44 – Frecuencias y dificultades para los ítems del caso 2 de IRALE. *220*
Tabla 45 – Cálculo detallado de las dificultades en los ítems ambiguos en el caso 2 de IRALE. *220*

Abreviaturas y acrónimos

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.

a – Índice o parámetro de discriminación del ítem.

ADL – Advanced Distributed Learning. Aprendizaje distribuido avanzado.

b – Índice o parámetro de dificultad del ítem.

BOGA – Versión online del sistema Hezinet.

c – Índice o parámetro de pseudoacierto del ítem.

CCI – Curva Característica del Ítem.

CEFR – Common European Framework of Reference for Languages. Marco europeo común de referencia para los idiomas.

C.ex – Criterio de aceptación/rechazo de un experto.

C.it – Criterio de aceptación/rechazo de un ítem.

DCMI – Dublin Core Metadata Initiative. Iniciativa de metadatos Dublin Core.

EGA – Euskararen Gaitasun Agiria. Certificado de conocimientos básicos de euskara.

ELSA – E-Learning Systems Architecture. Arquitectura para sistemas e-learning.

EOI – Escuela Oficial de Idiomas.

EVA – Espacio Virtual de Aprendizaje.

GHyM – Grupo de investigación de Hipermedia y Multimedia.

HABE – Helduen Alfabetatze eta Berreuskalduntzerako Erakundea. Institución para la alfabetización y euskaldunización de adultos.

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

HIZEBA – HIZkuntzaren EBAluazioa. Evaluación del idioma.

IC – Intervalo de Confianza.

IEA – Association for the Evaluation of Educational Achievement. Asociación para la evaluación del logro educacional.

IEEE – Institute of Electrical and Electronics Engineers. Instituto de ingenieros eléctricos y electrónicos.

IEEE-LOM – Learning Object Metadata del IEEE. Metadatos de objetos de aprendizaje del IEEE.

IMS – Instructional Management Systems. Sistemas de gestión instruccional.

IMS CC – IMS Common Cartridge. Estándar cartucho común del IMS.

IMS CP – IMS Content Packaging. Estándar empaquetado de contenidos del IMS.

IMS GLC – IMS Global Learning Consortium, Inc. Consorcio de aprendizaje global de IMS.

IMS QTI – IMS Question & Test Interoperability. Estándar para la interoperabilidad de ítems y test del IMS.

INES – International iNdicators of Education Systems. Indicadores internacionales de los sistemas de educación.

IRALE – IRakasleen ALfabetatze eta Euskalduntzea . Alfabetización y euskalduniación del profesorado.

LMS – Learning Management System. Sistema de gestión del aprendizaje.

LCMS – Learning Content Management System. Sistema de gestión de contenidos de aprendizaje.

MDDM – Métodos de Desarrollo Dirigidos por Modelos.

M.dif – Estadístico definido ad hoc en el experimento Hezinet para estimar el valor del parámetro dificultad del ítem empleando juicios de expertos.

MMANCA – MetaModelo de ANálisis y CÁlculos.

MMAP – MetaModelo de APortaciones.

MMCA – MetaModelo de CAlibraciones.

MMCU – MetaModelo de CUestionarios.

MMRE – MetaModelo de REsultados.

OA – Objeto de Aprendizaje.

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

p – Nivel de significación estadístico dificultad en la TCT. Proporción de individuos que acertaron un ítem entre el total que lo contestaron.

PC1 (*PC2*) – Prueba de Campo 1 (2) con expertos en el experimento Hezinet.

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.

Rpbis – Correlación biserial – puntual.

SCORM – Shareable Content Object Reference Model. Modelo de referencia de objetos de contenido compartibles.

SHA – Sistema Hipermedia Adaptativo.

SIETTE – Sistema Inteligente de Evaluación mediante Test para TeleEducación.

SQL – Structured Query Language. Lenguaje de consulta estructurado.

SweSAT – Swedish Scholastic Aptitude Test. Pruebas de admisión a las universidades de Suecia.

TAI – Test Adaptativo Informatizado.

TCT – Teoría Clásica de los Test.

θ – Nivel de habilidad o rasgo del alumno.

TIMSS – Trends in International Mathematics and Science Study. Estudio internacional de tendencias en matemáticas y ciencias.

TRI / IRT – Teoría de Respuesta al Ítem / Item Response Theory.

UML – Unified Modeling Language. Lenguaje de modelado unificado.

UPV/EHU – Universidad del País Vasco/Euskal Herriko Unibertsitatea.

WWF – Windows Workflow Foundation. Fundación de flujos de trabajo de Windows.

XML – eXtensible Markup Language. Lenguaje de marcado extensible.

XSD – XML Schema Definition. Definición de esquema XML.

Parte Primera: Introducción

En la Parte Primera, dedicada a presentar la **introducción** a la investigación desarrollada, se engloban los dos primeros capítulos de esta tesis doctoral.

El capítulo **I - Objetivos y contexto** describe brevemente la justificación y los objetivos de la tesis, junto con el trabajo realizado, y ubica el mismo en el contexto del grupo de investigación del que forma parte. En este capítulo también se presenta la guía de lectura de la presente memoria.

El capítulo **II - Antecedentes: Calibración de los ítems de Hezinet** presenta con detalle el pilar básico que ha servido como inspiración a esta tesis: los experimentos de calibración llevados a cabo sobre un banco de ítems para el sistema Hezinet y que utilizaron el juicio de expertos para estimar la dificultad de cada ítem.

I Objetivos y contexto

Todo sistema de aprendizaje debe complementarse con algún mecanismo de evaluación de los alumnos, que por lo general sirve para medir la adquisición del conocimiento que se ha producido tras un cierto grado de interacción con el sistema. En el caso concreto de los sistemas e-learning, esto es, de los entornos de aprendizaje que hacen uso de las tecnologías de información y comunicación, uno de los principales factores de éxito reside precisamente en la capacidad de estos mecanismos para evaluar eficazmente los conocimientos que está adquiriendo el alumno. El modo habitual de llevar a cabo esta evaluación consiste en la administración de test a los alumnos, ya sea utilizando el formato tradicional o bien empleando test adaptativos.

Los test adaptativos informatizados (TAI) son test que emulan el comportamiento de un evaluador humano, y que se generan dinámicamente basándose en las respuestas que va dando el examinado. La idea es que si el evaluado falla una pregunta, la siguiente cuestión que se le plantee será más fácil; y viceversa, como consecuencia de un acierto se administrará un ítem ligeramente más difícil. Los TAI ofrecen multitud de ventajas frente a los test tradicionales sobradamente discutidas en la literatura (Kingsbury y Weiss, 1983; Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg y Thissen, 2000), aunque para garantizar su buen funcionamiento es necesario estimar los parámetros del modelo de la Teoría de Respuesta al Ítem (TRI) – como mínimo, la dificultad – que caracterizan las preguntas. Este proceso de estimación se denomina *calibración* del banco de ítems y posee muchas utilidades al margen de los TAI, por ejemplo permite comparar resultados de un año con los de los años anteriores o siguientes. Sin embargo, en el contexto de esta memoria la calibración nace como una necesidad a la hora de implementar los TAI.

Históricamente los ítems se han calibrado estimando únicamente su dificultad siguiendo la experiencia de individuos doctos en la materia sobre la que tratan. Con la aparición de la TRI surge la calibración estadística (o psicométrica) que consiste en administrar los ítems a una muestra muy amplia de individuos para que los respondan y, a partir de sus respuestas, estimar estadísticamente los valores de los parámetros de la TRI que los caracterizan (generalmente dificultad, discriminación y pseudoacierto). Si se utiliza este último método, se debe abordar un procedimiento computacionalmente complicado que maneja una gran cantidad de datos y hace imprescindible el uso del ordenador, lo que constituye un problema a la hora de llevarlo a cabo ya que se requieren conocimientos multidisciplinares en psicometría, estadística e informática que las personas interesadas en realizar una calibración no tienen por qué dominar. Por este motivo, aún hoy en día los ítems se siguen calibrando mayoritariamente a partir de estimaciones de expertos (Muñiz, 2010).

En el contexto de esta investigación *calibrar un conjunto de ítems mediante el juicio de expertos* es un proceso que consiste en establecer en una métrica común la dificultad de cada ítem. Para ello se parte de una colección de preguntas o ítems que evalúan un determinado tema y de una muestra de datos que ha sido recogida

previamente a individuos especialistas en ese tema y que contiene, como mínimo, estimaciones sobre el nivel de dificultad de cada uno de ellos. Este proceso de calibración tampoco es una tarea simple, pues primero hay que comprobar la fiabilidad de los expertos a involucrar y asegurarse de la corrección de los ítems, es decir, de la no existencia de ítems anómalos. Asímismo, debe implementarse algún método para consensuar los criterios, a veces dispares, de los expertos consultados. Posteriormente, puede que sea preciso repartir el conjunto de ítems en distintos cuestionarios a administrar, siendo necesario incluso contar con que algunos ítems se repitan en varios cuestionarios. Tras ello, se realiza la recogida y análisis de todos los datos y cabe destacar que, aunque los métodos ideados para analizar las respuestas recogidas sean simples, el análisis manual de los datos no es fácil y en muchas ocasiones será preciso el uso del ordenador.

Este trabajo de tesis presenta el análisis del proceso de calibración mediante el juicio de expertos, junto con el diseño e implementación de un sistema inteligente de ayuda, denominado CALLIE, que automatiza dicho proceso y guía al responsable de la calibración en su toma de decisiones. CALLIE permite seleccionar entre calibración puramente basada en la TRI (o estadística) o calibración mediante el juicio de expertos. La información relativa a la primera se puede encontrar en la memoria de Armendariz (2014). La presente memoria se centra en la presentación del sistema y la herramienta para la realización del segundo método, esto es, la estimación de la dificultad de los ítems basada en el juicio de expertos.

La decisión de desarrollar una herramienta como CALLIE surgió al haber constatado que la calibración de un banco de ítems es un proceso largo y tedioso que requiere conocimientos en varias disciplinas, lo que supone una restricción particularmente difícil de cubrir para la mayoría de individuos potencialmente interesados en llevar a cabo tal proceso. Además, se comprobó que en la actualidad no existe otra herramienta software de características similares, ya que ésta unifica en una sola aplicación los dos métodos de calibración mencionados y cubre el proceso de una manera global, desde la selección de los ítems a calibrar hasta la consecución de los valores de los parámetros.

Una vez desarrollada la aplicación CALLIE, se han llevado a cabo una serie de pruebas que han demostrado la validez de los distintos componentes de la herramienta.

Aunque el sistema CALLIE ha sido diseñado especialmente para integrar y extenderse a distintas formas de realizar una calibración, actualmente solo cubre las dos variantes que se mencionan en los trabajos de investigación de la siguiente sección y que fueron las seguidas durante la calibración de un banco de ítems para la evaluación de la lengua vasca, dejando para versiones futuras cualquier otra alternativa existente.

I 1 Contexto del trabajo realizado

La presente tesis doctoral sigue la línea de investigación abierta por el Grupo de Hipermedia y Multimedia (GHyM) enclavado en el departamento de Lenguajes y Sistemas Informáticos de la Universidad del País Vasco (UPV/EHU), que comenzó a trabajar en el área del e-learning en 1994, con el diseño del sistema hipermedia

adaptativo Hypertutor (Gutiérrez, Pérez, Usandizaga y Lopistéguy, 1996). Después de realizar sobre él ciertas modificaciones, se obtuvo lo que a día de hoy es el estandarte del grupo: *Hezinet*, un Sistema Hipermedia Adaptativo (SHA) para el aprendizaje de la lengua vasca (*euskera*) a través de Internet, que ha sido implantado en más de 60 *euskaltegis* o centros de enseñanza de euskera para adultos en el País Vasco, así como en varias Casas Vascas repartidas por Europa y el continente americano.

Tras la presentación en el mundo académico de Hezinet (Pérez, 2000), se abrieron varias líneas de evolución que actualmente prolongan la existencia de este sistema. La primera de ellas es la más comercial y, por ello, la que menos tiene que ver con esta tesis, pero debido a la repercusión que ha tenido en los últimos años, conviene mencionarla. Y es que, desde que el sistema fuera adquirido por la institución para la alfabetización y euskaldunización de adultos (HABE) del Gobierno Vasco, Hezinet (bajo el nombre comercial de BOGA para su versión web) ha sido objeto de una fuerte campaña de marketing. Como consecuencia, el motor Hezinet-BOGA se ha implantado no sólo en euskaltegis del País Vasco, sino que además se ha instalado en algunos centros educativos y culturales ubicados en Cataluña, Madrid, Argentina, Chile, Estados Unidos, Francia, Inglaterra, México, Puerto Rico, Uruguay y Venezuela.

En 2002, con el objetivo de ir mejorando el sistema se decidió construir el sistema ELSA (Armendariz, López-Cuadrado, Tapias, Villamañe, Sanz-Lumbier y Sanz-Santamaría, 2003), basado en Hezinet, pero que permitía incorporar nuevas tecnologías, estándares e investigaciones que se estaban realizando en aquellos momentos. Entre otras líneas, se decidió trabajar en los mecanismos de evaluación de alumnos del SHA, que actualmente continúa y que corresponde a dos proyectos de análisis y mejora del sistema como entorno educativo, ambos culminados con la elaboración de sendas tesis doctorales.

La primera tesis "Evaluación mediante test adaptativos informatizados en el contexto de un sistema adaptativo para el aprendizaje de la lengua vasca" (López-Cuadrado, 2008) recoge el trabajo realizado en lo referente al test de ingreso, esto es, la evaluación que se efectúa a cada nuevo estudiante con el fin de determinar su nivel de inicio en Hezinet. Se profundizó en el uso de la TRI para implementar TAIs de ingreso, con el objetivo de establecer y formalizar el proceso de calibración del banco de ítems mediante un método basado en las técnicas estadísticas y en los principios de la TRI. Como producto de esta tesis se obtuvo la base para el primer SHA de aprendizaje del euskera que proporciona evaluación adaptativa del estudiante. Esta primera línea permitió dotar a Hezinet de un módulo de evaluación que genera test adaptativos informatizados de ingreso al sistema.

La segunda tesis "E-learning y la calibración de ítems de test: Teoría de Respuesta al Ítem versus calibración basada en juicios de expertos. Un estudio empírico" (Arruabarrena, 2010) ha buscado no sólo determinar la validez de una calibración de ítems realizada a partir de la información proporcionada por expertos, sino además compararla con los resultados obtenidos a partir de una calibración estadística. Para ello, planteó dos propuestas de proceso para la calibración de ítems, una utilizando la TRI y otra utilizando el juicio de expertos, y estableció ciertas métricas asociadas para evaluar el consumo de recursos que conllevan. Las dos propuestas enumeradas se pusieron en práctica para calibrar un mismo conjunto de ítems utilizando métodos síncronos y asíncronos para la recogida de datos en cada uno de los casos. Como resultado de la experiencia se formalizó un único procedimiento para la

calibración de ítems que puede instanciarse para ser utilizado tanto en una calibración mediante juicios de expertos como en la calibración estadística. Además, se compararon los resultados obtenidos tanto respecto al tipo de calibración como respecto a los recursos consumidos.

Durante el transcurso de estas investigaciones se renovó la arquitectura del sistema Hezinet dando lugar a una nueva arquitectura: el hiperentorno adaptativo de aprendizaje genérico ELSA (López-Cuadrado, Armendariz y Pérez, 2006). ELSA proporciona una infraestructura e-learning, basada en servicios web sobre la plataforma .Net de Microsoft, que cumple con los estándares internacionales *IMS Question & Test Interoperability v1.2* (Smythe, Shepherd, Brewer y Lay, 2002) y *ADL SCORM* (ADL, 2001) para representar el dominio didáctico. Se trata de una arquitectura orientada a la evaluación, que incorpora nuevos elementos con respecto a Hezinet (López-Cuadrado, Pérez, Sanz-Santamaría, Armendariz, Gutiérrez y Vadillo, 2007), entre los que cabe destacar la herramienta de autor ADISTI (López-Cuadrado, Armendariz y Pérez, 2003) para la edición y el almacenamiento de ítems, el módulo de evaluación y el módulo de calibración, que a pesar de los cambios sufridos, ya se mostraban en las primeras aproximaciones (López-Cuadrado, Armendariz, Pérez y Arruabarrena, 2008).

Por otro lado, las dos vías citadas de investigación requirieron la revisión previa (Arruabarrena, 2005) y calibración de un banco de ítems de evaluación para el test de ingreso en el sistema que han servido de base para crear el sistema de clasificación de alumnos de BOGA (www.ikasten.ikasbil.net). Se presentó un plan de evaluación específico para determinar la validez de una calibración de ítems utilizando estimaciones de expertos (Arruabarrena, Vadillo y Gutiérrez, 2003) frente a la calibración por medios estadísticos que propugna la TRI (López-Cuadrado, Pérez, Vadillo y Arruabarrena, 2002). Ajustándose a este plan de evaluación, se desarrollaron paralelamente ambas calibraciones de los ítems. Primero, un grupo de expertos estimó el parámetro dificultad de los ítems cumplimentando distintos cuestionarios en papel y después, tomando como base las estimaciones anteriores para diseñar los subtest a administrar, se realizó una calibración estadística para estimar los parámetros de dificultad, discriminación y pseudoacierto. Así, se consiguieron dos calibraciones del banco de ítems mediante dos métodos diferentes: mientras que la primera estableció el valor de la dificultad para cada ítem a partir de la información proporcionada por los expertos, la segunda de ellas les asignó los valores obtenidos estadísticamente a partir de las contribuciones de individuos no expertos.

Una vez definidos los procesos de calibración para los dos métodos (Arruabarrena, 2010; López-Cuadrado, Pérez, Vadillo y Gutiérrez, 2010) y realizados distintos estudios que analizaban y comparaban ambos procesos, sus costes y su complejidad (Arruabarrena, López-Cuadrado y Armendariz, 2007; Arruabarrena et al., 2003; López-Cuadrado, Armendariz, Pérez, Arruabarrena y Vadillo, 2009), se tuvo en cuenta que la necesidad de disponer de ítems calibrados está siempre latente en un sistema de aprendizaje, y se decidió definir la presente tesis para automatizar el proceso de calibración mediante el juicio de expertos junto a otra tesis paralela (Armendariz, 2014) que automatizase el proceso de calibración psicométrico. Así, se fijó el objetivo de diseñar e implementar una herramienta informática denominada CALLIE de modo que se integrara en Hezinet utilizando la arquitectura renovada ELSA y que permitiera al usuario calibrar un banco de ítems de cualquier materia por la vía que prefiera, esto es, utilizando valoraciones de expertos o por medio de la calibración psicométrica.

I 2 Requerimientos y objetivos

La herramienta de ayuda CALLIE se concibe para dar cobertura a los dos métodos de calibración, por lo que se ha decidido dividir su diseño e implementación en dos subsistemas complementarios que deben integrarse: CALLIE-TRI para las calibraciones basadas en la TRI (Armendariz, 2014) y CALLIE-EXPERT para las calibraciones basadas en el juicio de expertos, que es el objeto de esta memoria. Por este mismo motivo, a la hora de plantear el sistema de calibración CALLIE-EXPERT, se parte de dos tipos de requerimientos: por un lado, los que permiten la compatibilidad con Hezinet y la integración de los dos métodos en una única herramienta y por otro, los específicos para llevar a cabo la calibración del banco de ítems vía expertos. El cumplimiento de estos requisitos posibilitará que CALLIE sea adaptable a cualquier materia de aprendizaje y pueda reutilizarse en otros ámbitos y dominios.

Respecto al primer tipo de requerimientos, CALLIE surge como una mejora para el sistema Hezinet tras añadirle la capacidad de generación de TAIs, con lo que la herramienta debe ser compatible con las características de este sistema. Concretamente, con objeto de implementar el módulo didáctico de calibración se utilizará tecnología de Microsoft, el *formato IMS QTI* para los ítems a calibrar, y se seguirá la arquitectura *e-learning ELSA*. Esta arquitectura define tres componentes principales: la funcionalidad, la capa de datos y los usuarios del sistema, de modo que cada uno de ellos dará lugar a una serie de requerimientos que CALLIE deberá cumplir. Así, respecto a la funcionalidad, CALLIE deberá permitir calibrar bancos de ítems mediante expertos o técnicas psicométricas para su uso posterior, por ejemplo en los TAIs. En cuanto a la capa de datos, CALLIE deberá constar de uno o más repositorios en los que se guarde la información necesaria para soportar todo el proceso de calibración del banco de ítems y los propios ítems calibrados. Por último, CALLIE deberá considerar los distintos tipos de usuarios que pueden aparecer en el proceso independientemente del método de calibración a llevar a cabo. ELSA además determina la arquitectura física a implementar, por lo que CALLIE deberá contemplar un *modelo de software en tres capas*: interfaz de presentación, lógica de negocio y datos.

Siguiendo con este primer tipo de requerimientos, y respecto a la construcción del sistema CALLIE, cualquier experimento de calibración a llevar a cabo íntegramente mediante ordenador impone varias exigencias. Estas exigencias surgen de las tres tareas básicas que constituyen el proceso de calibración, de las que se profundizará en la sección VI 2: la de diseño del experimento, la de administración de ítems y la de análisis de datos y calibración. Para empezar, existen dos subsistemas que se corresponden con dos artefactos informáticos a diseñar y generar: el que administra los ítems a los individuos participantes y el que calcula los parámetros de cada ítem a partir de los datos recopilados con esas administraciones. Para esta labor de diseño, se propone el uso de *técnicas de ingeniería dirigida por modelos* por lo que se deben modelar las distintas características que puede presentar un determinado experimento de calibración. También se plantea el análisis y automatización de los principales procedimientos de cálculo, que en general son procesos con múltiples alternativas y –especialmente en las calibraciones estadísticas – pueden necesitar el apoyo de otras aplicaciones externas. En este caso, se considera que la mejor opción es servirse de la tecnología de *workflows* y recurrir a *servicios web* cuando CALLIE deba comunicarse

con esas otras aplicaciones. En base a lo expuesto, el sistema CALLIE se concibe como un sistema inteligente de ayuda que permite realizar calibraciones de ítems y que es capaz de: *conocer las características del experimento* que se tiene que realizar y ayudar al individuo que desea calibrar sus ítems en las decisiones relevantes para especificarlo*; generar y opcionalmente configurar* un artefacto informático para llevar a cabo el experimento; y *generar* un workflow que lleve a cabo los procedimientos de cálculo especificados para obtener la calibración.

En cuanto al segundo tipo de requerimientos, los específicos, se deben a las exigencias que impone la informatización completa del proceso de calibración vía expertos. Esta informatización en particular exige que, una vez diseñados los cuestionarios a partir de la colección de ítems, la herramienta automatice su administración a los expertos, es decir, CALLIE-EXPERT ha de generar automáticamente un sistema informático que permita autentificar a los distintos expertos participantes, repartirles los cuestionarios y recoger sus respuestas, para posteriormente filtrarlas y calibrar los ítems. Obviamente, una implementación desde cero que sea mínimamente eficiente es una labor muy compleja y costosa puesto que implica como poco la creación de una Web, la creación de test, la gestión de los expertos, el control de su participación, el envío de emails y avisos, etc. Así, al abordar la automatización de esta administración de cuestionarios a los expertos y con el objetivo de reducir la complejidad de la herramienta CALLIE-EXPERT, se toma como hipótesis de partida aprovechar, para ese cometido, algún otro sistema ya existente. Concretamente y considerando que los usuarios más probables de este tipo de herramienta serán profesores, se plantea utilizar una *plataforma educativa Web* de las muchas disponibles en el mercado y que sea popular en ese colectivo, e integrarla en CALLIE-EXPERT.

Para cumplir con los requerimientos mencionados, en la presente tesis se han planteado seis objetivos fundamentales:

(1) *Garantizar la integración* en una sola herramienta informática de los elementos que calibran ítems mediante el juicio de expertos y los de la calibración estadística, herramienta que además deberá ser compatible con el sistema Hezinet. Para ello, se utiliza como punto de partida la propuesta del proceso global de calibración de Arruabarrena (2010) válida para ambos tipos de calibración y, al igual que para la parte psicométrica, el prototipo teórico para CALLIE propuesto por López-Cuadrado, Armendariz et al. (2009), y el entorno de desarrollo MS Visual Studio con Visual Basic y ASP .NET como lenguajes principales de programación.

(2) *Diseñar e implementar el modelo de especificación*. Diseñar e implementar un modelo para la calibración de ítems vía expertos.

(3) *Diseñar e implementar el subsistema de administración de ítems*. Diseñar e implementar un sistema Web para la administración electrónica de cuestionarios a los expertos que sea compatible con el sistema Hezinet, utilizando para ello el estándar IMS QTI y una plataforma de e-learning compatible con este estándar.

(4) *Diseñar e implementar el subsistema de análisis y cálculos*. Diseñar e implementar el resto del proceso global para la calibración de ítems utilizando el juicio de expertos para estimar el parámetro dificultad como se propone en Arruabarrena (2010) e integrando el sistema de administración anterior.

(5) *Diseñar e implementar la herramienta final* de ayuda para calibraciones vía expertos, integrando los elementos anteriores y con una interfaz amigable que guíe en todo momento al responsable de la calibración durante el proceso.

(6) *Probar la validez del sistema* implementado para la calibración mediante el juicio de expertos.

I 3 Organización de la memoria y guía de lectura

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

La *primera parte: Introducción* consta de dos capítulos en los que se presenta el trabajo de la tesis en su contexto y los antecedentes de la investigación. El primero, éste, pretende definir y contextualizar el trabajo realizado. Por su parte, el capítulo II describe el pilar básico que ha servido como punto de partida, a saber, la metodología de trabajo de los experimentos llevados a cabo durante la calibración del banco de ítems de Hezinet utilizando el juicio de expertos para estimar la dificultad de cada ítem, junto a los resultados obtenidos y conclusiones extraídas.

La *segunda parte: Fundamentos* consta de cuatro capítulos en los que se detalla la base teórica subyacente al trabajo de investigación realizado en esta tesis. Esta base se centra en los conceptos relevantes tenidos en cuenta para llevar a cabo la calibración de bancos de ítems mediante el juicio de expertos, y que son, por un lado, los estándares e-learning para la representación de ítems, y por otro, las técnicas de psicometría básicas que se utilizan en una evaluación mediante TAI y los fundamentos sobre la calibración de los bancos de ítems y sobre el proceso a seguir. Así, el capítulo III está dedicado a presentar los fundamentos de la psicometría, junto a sus dos teorías predominantes: la TCT y la TRI. El capítulo IV de la memoria trata los estándares en el marco de los contenidos de aprendizaje online, incidiendo en el estándar IMS para el modelado e intercambio de ítems y test entre sistemas. El capítulo V versa acerca de los bancos de ítems calibrados y su obtención en el marco de la TRI. Por último, el capítulo VI cita varios aspectos teóricos basados en la ingeniería del software y que son aplicables al proceso de calibración de ítems vía expertos; acto seguido, se describe el proceso global de calibración y se detallan los pormenores de este proceso en las calibraciones que utilizan el juicio de expertos.

La *tercera parte: Estado del arte* está formada por dos capítulos y versa sobre el estado del arte del software en e-learning referido respectivamente a la administración y calibración de ítems y test que se han utilizado en el trabajo de investigación de esta tesis. Así, el capítulo VII trata sobre las aplicaciones Web actuales que soportan estándares e-learning y – tras una serie de definiciones para facilitar su comprensión al neófito en el tema – compila distintas plataformas e-learning que permiten la creación y gestión de contenidos de aprendizaje para la administración de ítems y test utilizando el estándar IMS o afines, para finalizar centrándose en la empleada para la construcción de CALLIE en su parte de calibración vía expertos: Moodle. El capítulo VIII de la memoria recopila referencias a diferentes programas informáticos que pueden resultar

de ayuda a la hora de calibrar un banco de ítems mediante el procedimiento basado en el juicio de expertos.

La *cuarta parte: La herramienta de ayuda CALLIE-EXPERT* consta de cuatro capítulos en los que se documenta el grueso del trabajo realizado para dar cobertura a los objetivos inicialmente planteados en la tesis, es decir, se pormenoriza la especificación y diseño del sistema de calibración y la implementación de la herramienta de ayuda a la calibración CALLIE-EXPERT centrada en el procedimiento de calibración vía expertos, seguido de la explicación de las pruebas realizadas para verificar su validez y, para finalizar, se enumeran las conclusiones de la investigación. En concreto, el bloque formado por los dos primeros capítulos de esta parte detalla el diseño del sistema y la herramienta CALLIE-EXPERT para calibración mediante expertos. En el capítulo IX se detalla la arquitectura y metamodelos que utiliza la herramienta para automatizar el proceso global de calibración de ítems mediante el juicio de expertos. El capítulo X relata cómo la herramienta controla y guía el transcurso del proceso global de calibración de ítems previamente descrito, a través de la explicación de la interfaz de los dos módulos principales de CALLIE-EXPERT: CALLIE-ESKARI y CALLIE-PRO. Por su parte, en el capítulo XI de la memoria se detallan las pruebas realizadas con la herramienta en dos casos reales de calibración vía expertos y se evalúan los resultados obtenidos mediante su utilización. Para finalizar, el capítulo XII 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 de la autora.

Por último, la *quinta parte: Anexos y Bibliografía* 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 primera), se asientan los fundamentos teóricos subyacentes (parte segunda) junto con el estado del arte (parte tercera) para documentar el trabajo realizado junto con las pruebas realizadas a la herramienta CALLIE y concluir (parte cuarta). 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 los estándares e-learning IMS, el capítulo IV puede no ser relevante; quien tenga conocimientos de psicometría y calibración puede descartar la lectura de los capítulos III y V; el experto en plataformas e-learning LMS puede prescindir de la sección VII 1. Por otro lado, quien solo desee tener un conocimiento superficial de los resultados de este trabajo, puede centrar su lectura en los capítulos I , XI y XII junto con la sección IX 3; por último, si está interesado en la integración de la herramienta CALLIE con la plataforma educativa Moodle debería leer el capítulo IX junto con la sección VII 2.

II Antecedentes: Calibración de los ítems de Hezinet

Este capítulo muestra el trabajo realizado en la calibración mediante juicio de expertos de un banco de 252 ítems de selección múltiple en formato de texto para el TAI de ingreso al sistema Hezinet (Arruabarrena, 2010). Estos ítems de partida fueron elaborados y entregados por los productores de contenidos de la fundación cultural Aurten Bai/Zornotzako Barnetegia, y todos incluían un enunciado corto de tipo texto y cuatro posibles respuestas, de las que sólo una era correcta.

La calibración consistió en realizar un experimento en el que se administraron los ítems a múltiples expertos con el objetivo de estimar para cada ítem dos parámetros: su *dificultad* y su *destreza*. En este capítulo se resumen los hitos y conclusiones más relevantes. El proceso completo, así como la enumeración de los datos manejados, se puede encontrar en (Arruabarrena y Pérez, 2005a), (Arruabarrena y Pérez, 2005b) y (Arruabarrena y Armendariz, 2008).

El capítulo se organiza de la siguiente manera: En primer lugar se comenta el diseño del experimento realizado. Después se describe el diseño de los cuestionarios asociado. A continuación se indican las pruebas de campo planificadas, y para terminar, se describe el análisis de datos que se realizó, y que dio lugar a la calibración del banco de ítems.

II 1 Diseño del experimento

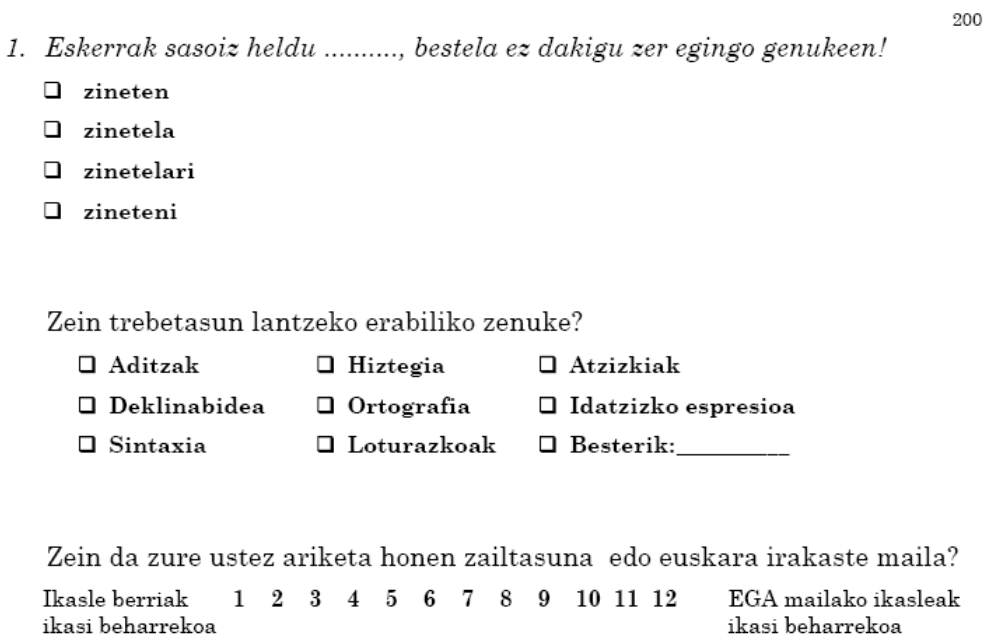
El objetivo del experimento *es recopilar 7 valoraciones diferentes de cada uno de los ítems del banco por parte de expertos voluntarios y sin remuneración* mediante encuestas plasmadas en cuestionarios de papel. Esta característica hace que sea crítico el diseño de los cuestionarios, que se decide que no exceda de 45 minutos. Se plantearon dos pruebas de campo: PC1 y PC2.

Los **sujetos activos** se centraron en 4 roles: una *desarrolladora/responsable principal* que se encargó de la coordinación y ejecución del proceso de calibración; un *supervisor*, que asumió la labor de controlar puntualmente el desarrollo de todo el proceso realizado; una *colaboradora*, que se ocupó de la grabación de los cuestionarios completados y de la elaboración de varios entregables; y el *responsable del proceso de calibración estadística* de los ítems, con el que hubo que coordinarse en alguna tarea compartida.

Los **sujetos pasivos** se enfocan en dos roles: el de *revisor* y el de *experto*. Los revisores eran filólogos o lingüistas de la lengua vasca con experiencia en el desarrollo

y estudio académico del euskara en la UPV-EHU y su labor consistió en detectar fallos en los ítems o en los cuestionarios, determinando si el documento que recibiría el experto tenía una estructura y tamaño razonables, si las instrucciones de cumplimentación eran claras, los ejemplos aclaratorios y si el apartado de valoración era cómodo de rellenar. Los expertos fueron profesores de euskera de euskaltegis, normalmente personas que trabajaban con euskera batua y que, posiblemente habían tenido la posibilidad de haber trabajado con el sistema Hezinet. Su cometido fue cumplimentar los cuestionarios diseñados.

El **diseño de los cuestionarios de las pruebas** incluía una portada en la que se indicaba el número de cuestionario, se presentaba el objetivo de trabajo y agradecía la participación voluntaria en el mismo, detallaba la forma de contacto con los responsables de la calibración y el modo de envío de los cuestionarios. A continuación, se presentaban las instrucciones de cumplimentación del cuestionario ilustradas con ejemplos concretos. En el resto del cuestionario se presentaban los datos que el experto debía cumplimentar. Un primer apartado de *datos personales* solicitaba datos no identificativos del experto. Posteriormente se presentaban los *ítems a valorar:* un subconjunto del banco de ítems. En estas pruebas se solicitaron a los expertos tres valoraciones por cada uno de los ítems (Figura 1): la *respuesta correcta*, la *destreza lingüística* a la que pertenecía de entre 8 habilidades lingüísticas y el *nivel de dificultad* utilizando la escala de 12 niveles contemplada por el currículo de HABE (1984). La primera sirvió como elemento de control, mientras que las otras dos constituyeron las dimensiones que se pretendían estimar con el experimento. Finalmente se solicitaban *aportaciones propias* al participante sobre todo el cuestionario en general.



200

1. *Eskerrak sasoiz heldu ....., bestela ez dakigu zer egingo genukeen!*
   - ❑ **zineten**
   - ❑ **zinetela**
   - ❑ **zinetelari**
   - ❑ **zineteni**

Zein trebetasun lantzeko erabiliko zenuke?

❑ **Aditzak**	❑ **Hiztegia**	❑ **Atzizkiak**
❑ **Deklinabidea**	❑ **Ortografia**	❑ **Idatzizko espresioa**
❑ **Sintaxia**	❑ **Loturazkoak**	❑ **Besterik:_____**

Zein da zure ustez ariketa honen zailtasuna edo euskara irakaste maila?

Ikasle berriak ikasi beharrekoa **1 2 3 4 5 6 7 8 9 10 11 12** EGA mailako ikasleak ikasi beharrekoa

Figura 1 – Detalle de un ítem del apartado del cuestionario *Ítems a valorar*.

Para la prueba de campo *PC1* se elaboraron 8 cuestionarios pidiendo aportaciones sobre 42 ítems cada uno. En estos cuestionarios se aplicó un diseño de anclaje, que consiste en incluir en todos los cuestionarios un mismo subgrupo de ítems. Así, de los 42 ítems de cada cuestionario, 12 eran comunes a todos y conformaron los

ítems de anclaje para la prueba. Para la prueba de campo *PC2* se formaron 6 cuestionarios de 42 ítems todos diferentes, esto es, sin ítems de anclaje.

Los cuestionarios confeccionados se pasaron a los revisores que valoraron positivamente los ejemplos incorporados en las instrucciones y consideraron adecuada la *estimación temporal de finalización* de 45 minutos. Además, analizaron los ítems de su cuestionario e indicaron algunas sugerencias de corrección que resultaron en que previamente se hiciera una revisión exhaustiva del banco de ítems original (Arruabarrena, 2005).

II 2 Administración de los ítems

En la *prueba de campo PC1* se recogieron 74 de los 80 cuestionarios que se precisaban, que contenían 3119 valoraciones de ítems. Durante la *prueba de campo PC2*, se recogieron 42 cuestionarios. Toda la información aportada por los cuestionarios se transcribió y almacenó en una base de datos. El tamaño de la muestra recogida fue de *4887 entradas/aportaciones* que corresponden a *116 expertos* participantes y todas realizadas sobre el banco de *252 ítems*.

II 3 Análisis de datos y calibración

Para analizar los datos se definieron varios criterios de depuración para aportaciones erróneas y/o anómalas de la muestra recogida en dos sentidos: *ítems* y *expertos*. Estos criterios se aplicarían de forma combinada hasta obtener una muestra estable (Arruabarrena y Armendariz, 2008).

En primer lugar, se establecieron dos familias de criterios basadas en estas ideas: la *familia de criterios C.ex* y la *familia de criterios C.it*. Estos cuatro criterios se enunciaron y detallaron como sigue (Arruabarrena, 2010):

- *C.ex-1. Análisis Aportación: Las aportaciones de expertos sobre los ítems se consideran válidas siempre que indiquen solo un nivel válido de dificultad*. C.ex-1 elimina cada aportación que no tenga estimación de nivel, tenga más de una o no esté dentro de los niveles válidos.
- *C.ex-2. Análisis Aciertos: Se eliminan los cuestionarios de expertos que no superan un porcentaje de respuestas correctas.* Se consideró no fiable la administración de aquellos expertos que no llegaron a un mínimo de acierto del 75% de los ítems válidos respondidos.
- *C.it-1. Análisis Aciertos Item: Un ítem se acepta si un porcentaje mínimo de los expertos responde correctamente al mismo.* Este criterio *C.it-1* obtiene para cada ítem su porcentaje de aciertos, y lo considera no fiable si su tasa de acierto es inferior a un umbral dado. Si este es el caso, se rechaza el ítem

junto con todas sus valoraciones. Se consideró no fiable todo ítem que no llegó a un mínimo de acierto del 70% por parte de los expertos.

- *C.it-2. Análisis Dispersión Item: Un ítem se mantiene si un porcentaje mínimo de las valoraciones de nivel dadas por los expertos para ese ítem se encuentran agrupadas en una horquilla determinada de niveles consecutivos de dificultad.* Se recomendó que este porcentaje varíe en un rango entre el 70% y el 85% y preferiblemente utilizar la malla más restrictiva, esto es, la eliminación del banco de aquellos ítems que no concentren al menos el 85% de opiniones válidas en un rango continuo correspondiente a un tercio (35%) de la escala numérica de dificultad.

En segundo lugar, para determinar la aplicación combinada de estos cuatro criterios en el experimento de Hezinet se decidió depurar la muestra de partida mediante la aplicación del algoritmo de la Figura 2.

Paso 1. Aplicar el criterio C.ex-1.

Paso 2. Aplicar el criterio C.it-1 con una tasa del 50%.

Paso 3. Aplicar de manera iterativa los otros dos filtros en el orden C.it-2 (con una horquilla de 4 niveles y un mínimo de valoraciones en ella del 75%) y C.ex-2 (con un umbral de acierto del 75%) hasta estabilizar los resultados.

Figura 2 – Algoritmo de aplicación de los filtros en el experimento Hezinet.

Tras este análisis de datos, la muestra final depurada quedó representada por 3315 aportaciones de 192 ítems realizadas por 111 expertos.

Respecto al cálculo de la dificultad final de cada ítem no retirado del banco se tuvo en cuenta que los expertos aportaban información subjetiva que no todos compartían y se fomentó buscar el consenso de sus juicios, para lo cual se ideó un procedimiento estadístico ad-hoc denominado M.dif (Arruabarrena, 2005).

M.dif se define mediante dos reglas que guardan relación con la criba de ítems y que establecen – a partir de juicios de dificultad emitidos por expertos – el valor más probable entre los pronósticos de dificultad más consensuados. La primera regla *M.dif-1* descarta los juicios más extremos de cada ítem, mientras que la segunda *M.dif-2* sirve para desambiguar cuando existan ítems comprendidos en dos intervalos contiguos, con el mismo número de niveles y la misma tasa de frecuencias de pronósticos dificultad. El enunciado de estas dos reglas es:

- *M.dif-1. Cálculo Dificultad.* La dificultad del ítem es el promedio de las frecuencias relativas de las valoraciones contenidas en el intervalo contiguo de X niveles (siendo X un tercio de la escala) con mayor densidad de valoraciones.
- *M.dif-2. Cálculo Ambigüedad.* Si hubiera más de un intervalo que cumpla la condición anterior, entonces se extenderá el intervalo con un nivel más y se escogerá el intervalo con X+1 niveles consecutivos con más valoraciones y menor desviación.

Estos cálculos redujeron las valoraciones a 2933 de los 192 ítems estudiados realizadas por los 111 expertos con los que se contaba.

La aplicación de M.dif dejó patente que la distribución de las dificultades de los ítems del banco era desigual: la mitad del banco tenía una dificultad intermedia, y el resto estaba en torno a ésta, tendiendo hacia niveles básicos. Igualmente quedó patente la escasez de ítems con estimaciones de dificultad elevada. Esta distribución de las estimaciones calculadas era la esperada, a pesar de no estar distribuidas uniformemente a lo largo de la escala de dificultad [1,12], ya que el banco de ítems se venía empleando en un euskaltegi de gran envergadura para determinar el nivel de entrada de nuevos alumnos, y el gran bloque de alumnos que ingresan lo hacen en los niveles intermedios e iniciales, y en ese orden. Además, *los resultados fueron coherentes con los obtenidos en el proceso de calibración estadística* que se realizó en paralelo (López-Cuadrado, 2008).

Finalmente se llevó a cabo un *estudio de funcionamiento diferencial* de los ítems comparando los resultados de las dos pruebas de campo realizadas con objeto de determinar posibles diferencias significativas entre ambos tipos de pruebas. Para ello, se estudiaron las aportaciones de los expertos desde el punto de vista del cribado de la muestra y de los juicios emitidos por los expertos. Los cálculos se desglosaron por cada prueba de campo con el fin de concluir si hubo o no diferencia funcional entre los pronósticos de los expertos de la PC1, los de la PC2 y todo el conjunto de pronósticos.

Con respecto al cribado de la muestra, con objeto de estudiar *la evolución de las aportaciones recopiladas* en la PC1 y en la PC2, se desglosó por prueba de campo y por ítem el número de aportaciones de la muestra total en tres momentos: al finalizar la recogida de la información, al finalizar el primer cribado de la muestra y al estimar la dificultad de los ítems. En cada uno de estos momentos se consideraron dos variantes: incluir solo las aportaciones con respuesta al ítem acertada e incluir toda respuesta –correcta o no – dada al propio ítem por el experto. *Los resultados indicaron que se mantuvieron invariables tanto las proporciones de descarte de aportaciones como las de respuestas acertadas en los momentos considerados*.

Con respecto a los juicios o pronósticos de dificultad emitidos por los expertos, se quiso comprobar si la dificultad estimada con la muestra que contenía tanto aportaciones de la PC1 como de la PC2, – a la que se denominó PC1&2 –, coincidía con las dificultades estimadas únicamente con los valores de la PC1 o con los de la PC2 o con los de ambas, *para corroborar en qué medida las dificultades estimadas eran pronósticos consensuados por los expertos* de las diferentes muestras. Se concluyó que *hubo concordancia entre los pronósticos* emitidos por los expertos de la PC1, los expertos de la PC2 y todo el conjunto de expertos. Considerando que la dificultad estimada se calculó para concretar el valor de dificultad consensuado más probable, podría decirse que *las dificultades estimadas por ítem estaban consensuadas* en PC1, PC2 y en PC1&2.

Al estudiar separadamente las aportaciones de las dos pruebas de campo, PC1 y PC2, y a la vista de los resultados obtenidos, en cuanto a evolución de volúmenes de descarte de aportaciones y dificultades estimadas, no hubo diferencia significativa. Esta similitud de resultados puso de manifiesto la *corrección de la selección de la muestra aleatoria de expertos* y *refrendó la propia acreditación de expertos*. Los resultados también permitieron concluir que *no hace falta utilizar un subconjunto de ítems comunes en los distintos cuestionarios* (diseño de anclaje en PC1 frente a no anclaje en

PC2) *y que utilizando 7 valoraciones por ítem* (en PC2) *se alcanzan resultados prácticamente idénticos a cuando se utilizan más* (12 en PC1).

Parte Segunda: Fundamentos

La Parte Segunda está dedicada a presentar los **fundamentos** en los que se sustenta la implementación del proceso de calibración de un banco de ítems mediante el juicio de expertos.

El capítulo **III - Psicometría básica** presenta y describe el marco teórico psicométrico para la realización de una herramienta que calibra bancos de ítems incidiendo en aquellos aspectos a utilizar en una calibración con contribuciones de expertos.

El capítulo **IV - Estandarización de recursos educativos** define conceptos fundamentales sobre la estandarización de contenidos en aprendizaje electrónico (e-learning) centrándose en las especificaciones IMS para la representación e intercambio de los recursos de evaluación: ítems y test.

El capítulo **V - Bancos de ítems calibrados** versa acerca de las aplicaciones de los bancos de ítems calibrados y su obtención en el marco de la TRI en el contexto de los test de evaluación.

El capítulo **VI - Proceso de calibración mediante el juicio de expertos** cita varios aspectos teóricos basados en la ingeniería del software que son aplicables al proceso de calibración de ítems vía expertos. Se describe el proceso global de calibración y se detallan los pormenores de este proceso en las calibraciones mediante el juicio de expertos.

III Psicometría básica

En todas las ciencias es necesario cuantificar ciertas propiedades, pero en psicología se trabaja con características para las que no existen herramientas o reglas de medición directa, tales como la cantidad de conocimiento adquirido por un alumno, el nivel de ansiedad, el sentimiento de inferioridad, o la habilidad de una persona para desempeñar una determinada tarea. Por ello, para medir los rasgos psicológicos se recurre a modelos matemáticos o estadísticos que permiten realizar una estimación de la característica concreta a partir del rendimiento observado, englobados en una ciencia llamada *psicometría*, y que la Real Academia Española define como "*medida de los fenómenos psíquicos*".

En este capítulo se presenta el test como herramienta de medición de la habilidad o rasgo psicológico de una persona (sección III 1). Posteriormente se explican las dos grandes teorías psicométricas (sección III 2): la Teoría Clásica de los Test (TCT), el primer modelo formal utilizado para evaluar los rasgos psicológicos de los sujetos, y la Teoría de Respuesta al Ítem (TRI).

III 1 El test como medidor de rasgo o habilidad

Cuando se administra un test de evaluación, el objetivo es medir a un individuo con la idea de cuantificar una variable denominada genéricamente *rasgo*. Puede tratarse de la capacidad para retener y/o aplicar el conocimiento adquirido (general o de un tema específico), memoria, comprensión o cualquier otro tipo de habilidad. En el ámbito de la psicometría es habitual utilizar el término *habilidad* como sinónimo de rasgo (Baker, 2001; Birnbaum, 1968; López Pina, 1995) a pesar de no tener exactamente el mismo significado en algunos ámbitos como por ejemplo en los test de personalidad, que evalúan rasgos que no son habilidades. Sin embargo en esta memoria no se va a trabajar en estos ámbitos, por lo que se utilizarán ambas palabras para hacer referencia a la *variable que trata de cuantificar un test en los sujetos a los que les es administrado*.

Es complicado medir el rasgo o habilidad y asignarle un valor concreto, debido a que los test evalúan conceptos y no una dimensión física. A pesar de ello, cualquiera que sea la habilidad que mida un test, ésta se tendrá que poder determinar como un número real del intervalo $(-\infty, +\infty)$ cuyo punto central es el cero. Esta escala se ha establecido de esta forma porque simplifica mucho el trabajo. La idea principal es determinar un valor numérico para cada sujeto, que en el ámbito de la TCT se representará como V (puntuación verdadera) y en la TRI con la letra griega theta (θ), y que representa la habilidad de dicho sujeto. Así, resulta muy sencillo comparar las habilidades de varios individuos.

III 2 Teorías psicométricas

Una teoría psicométrica proporciona métodos para la construcción de test y provee de modelos matemáticos que facilitan la interpretación y validación de los resultados obtenidos. En concreto, uno de los aspectos más importantes que maneja una teoría psicométrica es el tratamiento de los errores de medida en lo referente a tres aspectos: minimización del error cometido por un test al estimar la habilidad de los examinados, correlación entre las variables y suministro de puntuaciones o estimaciones de habilidad con relación a un determinado nivel de confianza. Históricamente han destacado dos grandes teorías: la *Teoría Clásica de los Test* y la *Teoría de Respuesta al Ítem*, que se tratarán en los siguientes apartados.

III 2.1 La Teoría Clásica de los Test (TCT)

La Teoría Clásica de los Test, también conocida como *modelo clásico de la puntuación* o *teoría del error de medida*, surgió a partir del trabajo de Spearman (1904, 1907, 1913), aunque no recibió su forma axiomática hasta mediados los años sesenta (Novick, 1966). Esta teoría se fundamenta en el denominado *modelo clásico o lineal*, que establece una relación (lineal, de ahí su nombre) entre la habilidad del examinado y la puntuación obtenida en el test (Ecuación 1).

$$X = V + e$$

Ecuación 1 – Relación lineal entre puntuación real y observada.

Este modelo se expresa en términos de la puntuación empírica, el elemento central alrededor del cual gira toda la teoría. Concretamente, se considera que la *puntuación empírica* del sujeto (X), esto es, el valor observado en el test, es igual a la suma de dos componentes hipotéticos y desconocidos a priori: la *puntuación verdadera* o habilidad real del evaluado (V) y un determinado *error de medida* (e). Dado que para cada examinado y test existen dos incógnitas (V y e), se deben realizar ciertas suposiciones para poder resolver la ecuación anterior. Así, la TCT asume que el error es una perturbación aleatoria de la medición compuesta por diferentes factores (propios del sujeto, del test o externos), y supone que la media de su valor esperado es cero en la población de referencia y que además es independiente de la puntuación verdadera del examinado y de los errores de otros test.

Tras aplicar una prueba, lo único que se conoce es la puntuación empírica de los examinados, si bien de la relación entre X y V nada se sabe. Podría pensarse en correlacionar ambas variables para un grupo de personas, pero el problema seguiría siendo el mismo, a saber, que las puntuaciones verdaderas son desconocidas. Para solventar este inconveniente, lo que se hace es analizar la correspondencia entre las puntuaciones empíricas obtenidas por los sujetos que han realizado dos (o más) formas paralelas de un mismo test. Para la TCT el concepto de *formas paralelas* se refiere a diferentes test que evalúan el mismo rasgo, de tal manera que cada sujeto tiene la misma puntuación V en ambas formas, y la varianza de los errores de medida también es idéntica sea cual sea la forma del test utilizada para realizar la medición.

El *análisis clásico del ítem* está interesado en medir la dificultad y la discriminación del ítem, que se utilizarán como base para determinar si un ítem es bueno. El *índice de dificultad* del ítem (*P*) para la TCT es simplemente la proporción de individuos que contestaron correctamente al ítem entre el total de examinados que lo respondieron. El *índice de homogeneidad o discriminación*, indica en qué medida el ítem representa lo mismo que el test, y se evalúa mediante un coeficiente de correlación entre las puntuaciones obtenidas en el ítem (puntuadas con 0/1) y la puntuación total del test (correlación biserial-puntual, *rpbis*). Así, *rpbis* es una medida del poder discriminativo del ítem. Un valor positivo indica que los individuos que lo acertaron puntuaron alto, lo que generalmente significa que el ítem es bueno. En cambio, si el *rpbis* es negativo significa que los sujetos que puntuaron bajo respondieron mejor a este ítem que los que puntuaron alto, y el ítem necesitaría revisarse. Otro indicador para saber si un ítem es bueno es que la *media de individuos* que lo respondieron correctamente sea mayor que la de los que lo respondieron incorrectamente. Estas medias se calculan a partir del número de respuestas correctas y del número de respuestas incorrectas.

Además de evaluar cada ítem como un todo, las estadísticas pueden utilizarse para evaluar las opciones individuales de los ítems mediante un *análisis de distractores*. Lo que hace útil a las *estadísticas sobre las opciones* es la evaluación de las respuestas incorrectas, conocidas como *distractores*. En este caso, se calculan las *frecuencias de respuesta* por ítem y opción. Para la TCT, si la mayor parte de los individuos eligieron la opción correcta el ítem es fácil, si pocos la eligieron el ítem es difícil. Sin embargo, aunque pocos seleccionen la opción correcta esto no significa necesariamente que un ítem sea malo. Para saberlo, se examinan las *puntuaciones medias de respuesta* por ítem y opción si los que eligieron la opción correcta obtuvieron buenas puntuaciones y los que eligieron las demás obtuvieron puntuaciones relativamente bajas, entonces el ítem es todavía un buen ítem. En este tipo de análisis, un ítem malo tendría un *rpbis* negativo pero una dificultad moderada, lo que significaría que los examinados más hábiles seleccionaron más veces uno de los distractores que la respuesta correcta. Esto indicaría que el distractor en cuestión es posiblemente incorrecto o que el enunciado del ítem no está claro.

De la misma forma que se puede hacer un análisis del ítem se puede realizar un *análisis clásico del test*. Este análisis se basa principalmente en los valores obtenidos para cada ítem, con los que se calculan estadísticas descriptivas tales como la media, desviación típica, varianza, mínimo y máximo. La TCT también proporciona teoremas que permiten obtener empíricamente para un test medidas de fiabilidad, como por ejemplo el coeficiente de fiabilidad (que se define como la correlación entre los resultados de dos formas paralelas, y que viene a ser el cociente entre la varianza de las *V* y la varianza de las *X*, esto es, la proporción de varianza total que no se debe a errores) o el error típico de medida, y de validez, tales como el *coeficiente de validez* o el *error típico de estimación*.

La TCT presenta el inconveniente de que los supuestos que plantea no pueden contrastarse en la práctica (Muñiz, 2000). Concretamente, el primer problema dentro del marco clásico de la TCT es que las mediciones no resultan invariantes respecto al instrumento (test) utilizado. Para hacerlo se transforman las puntuaciones directas de los test en otras baremadas, por ejemplo en percentiles, con lo que se considera que se pueden ya comparar, y de hecho así se hace, lo que implica que se asume que los grupos

normativos en los que se elaboraron los baremos de los distintos test son equiparables, cosa que no tiene por qué ser cierta en la práctica. La segunda cuestión es la ausencia de invarianza de las propiedades de los test respecto de las personas utilizadas para estimarlas, es decir, propiedades psicométricas importantes de los test, tales como la dificultad de los ítems, o la fiabilidad del test, están en función del tipo de personas utilizadas para calcularlas, lo cual resulta inadmisible desde el punto de vista de una medición rigurosa. Por ejemplo, la dificultad de los ítems, o los coeficientes de fiabilidad dependen en gran medida de la muestra utilizada para calcularlos.

III 2.2 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 mencionadas en la TCT, y cuyas peculiaridades proporcionan un modelo teórico excelente para la elaboración de TAIs. 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 y ordenadores potentes. 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 (Lord y Novick, 1968), esto es, los supuestos y modelos más empleados. Las personas interesadas en ampliar conocimientos pueden consultar, entre otros, los libros de Hambleton y Swaminathan (1985), Hambleton, Swaminathan y Rogers (1991a), Lord (1980) y Muñiz (1996).

La TRI intenta dar una fundamentación probabilística al problema de la medición de rasgos inobservables. Para ello, 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 en función de su habilidad.

Esta teoría proporciona una familia de modelos matemáticos que se sustentan en el cumplimiento de diversos requisitos teóricos que deben cumplir los datos y que se asume los verifican. En este párrafo se enuncian los **supuestos de la TRI** o condiciones que se deben verificar cuando se usa la TRI como modelo de examen de comportamiento (Hambleton et al., 1991a; 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 postula 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 test (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. 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 test 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).

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. Así, 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 matemática monótona creciente denominada *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 $P(\theta)$ de que éste responda correctamente al ítem (eje vertical) (Figura 3). 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.

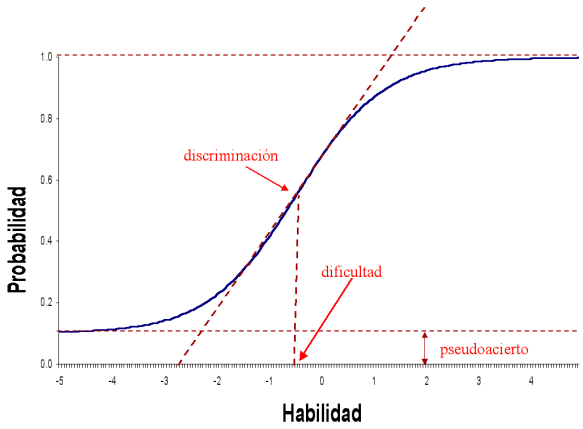


Figura 3 – Ejemplo de CCI para un modelo logístico de tres parámetros.

Los modelos se pueden clasificar en modelos *paramétricos*, aquéllos 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 en la práctica, siendo entre ellas la más ampliamente utilizada la familia de *curvas logísticas de uno (1PL), dos (2PL) o tres (3PL) parámetros* (Birnbaum, 1968). Estos modelos, en los cuales se centra esta memoria, son además modelos *dicotómicos* en los que se establecen para cada ítem sólo dos valores de respuesta posibles (correcta e incorrecta).

Dichos modelos de la TRI suponen que cada ítem tiene tres características destacables, tal y como ilustra la Figura 3, que se corresponden con los tres parámetros que los definen:

- Parámetro a_i o *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 o *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 $(-\infty,+\infty)$, 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 o *pseudoacierto*. Expresa la probabilidad de acertar un ítem cuando se desconoce la respuesta correcta, es decir, la probabilidad de acertar al azar.

El modelo 3PL – que considera la dificultad, la discriminación y el pseudoacierto del ítem – es el modelo utilizado habitualmente en las calibraciones estadísticas. Por su parte, el modelo 1PL constituye una particularización del modelo 3PL en el que solamente se considera la dificultad del ítem y se asumen valores constantes para los otros dos parámetros. Como lo único que es factible que un experto estime es precisamente la dificultad y esta tesis se centra en la calibración de ítems mediante el juicio de expertos, una calibración de este tipo basada en la TRI sólo tiene sentido a nivel práctico si se va a utilizar alguno de los modelos 1PL. Ambos modelos se detallarán en el resto de este apartado.

El **modelo de un parámetro** o 1PL, es el más simple, y supone que cada ítem tiene una única característica destacable: la *dificultad* (parámetro b_i), que comparte escala con la habilidad, y se define como el valor de θ para el que la probabilidad de una respuesta correcta es, en media, de 0,5. Cuanto mayor sea la dificultad del ítem, más cerca de $+\infty$ estará el punto de inflexión de su curva característica asociada; y viceversa: un valor pequeño de b_i indica que una persona con poca habilidad es capaz de responder correctamente el ítem con un 50% de posibilidades.

La habilidad del examinado y la dificultad del ítem se miden en la misma escala, cuyo origen y unidad son desconocidos. Para evitar arbitrariedades, cualquiera que sea el número de parámetros del modelo logístico utilizado, por regla general la media se sitúa en el cero, y la desviación típica se fija con valor 1 (Wainer y Mislevy, 2000). Si se tiene en cuenta que el 99,9% de los casos de la población se sitúa entre ±3 veces la desviación típica, los valores de las variables latentes b_i y θ difícilmente excederán el intervalo (-3,0, +3,0) (López Pina, 1995).

La Ecuación 2 presenta la expresión matemática de las CCI según el modelo logístico de un parámetro. En ella, $P(\theta)$ es la probabilidad de responder correctamente el ítem (cuya dificultad es b_i) cuando se tiene una habilidad θ. Por su parte, la constante D que aparece en la ecuación del modelo de un parámetro es un factor de escala; cuando es igual a 1 el modelo se corresponde con su versión logística tradicional, conocida como *modelo de Rasch* (1960), mientras que cuando D= 1,702 los valores obtenidos aproximan, para cualquier valor del continuo de habilidad, con un error inferior a la centésima a los de la curva normal acumulada de un parámetro (Haley, 1952).

$$P(u_i = 1 | \theta) = \frac{1}{1 + e^{-D(\theta - b_i)}}$$

Ecuación 2 – Modelo logístico de un parámetro.

El **modelo de tres parámetros** o 3PL es el modelo logístico más complejo con aplicación práctica, y supone que cada ítem tiene tres características destacables: *discriminación*, *dificultad* y *pseudoacierto* que se corresponden con los tres parámetros que definen el modelo. Así, las curvas logísticas del modelo 3PL se definen mediante la siguiente ecuación:

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

Ecuación 3 – Modelo logístico de tres parámetros (Birnbaum, 1968).

donde u_i=1 indica la probabilidad de que el sujeto responda correctamente el ítem i; y u_i=0 indica la posibilidad de respuesta incorrecta por parte del sujeto. La variable θ es el nivel de habilidad del examinado, es decir, lo que se mide mediante el test y, como ya se ha dicho, por lo general toma un valor real del intervalo [-3, 3]. Luego, $P(u_i=1|\theta)$ es la probabilidad de que un examinado elegido al azar con habilidad θ conteste correctamente al ítem i y la probabilidad de que falle ese ítem vendrá dada por $P(u_i=0|\theta) = 1 - P(u_i=1|\theta)$. La constante e es la base de los logaritmos neperianos, es decir, 2,71828. Según el modelo, los individuos con bajas habilidades (<0,0) tienen escasa probabilidad, en contraposición a los que tienen habilidades elevadas. La Ecuación 3 se emplea para modelar el modelo 3PL, pero también es una generalización de las ecuaciones correspondientes a los modelos 2PL y 1PL. De esta forma, 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 además 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.

Llegados a este punto, hay que recordar que los modelos de TRI son modelos muy restrictivos, 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.

III 2.3 Diferencias entre TCT y TRI

Las dos teorías conviven perfectamente en la construcción y análisis de ítems y test, puesto que cada una es aconsejable en determinadas situaciones. Tanto en la TCT como en la TRI, los valores de los parámetros de descripción de los ítems se calculan a partir de los datos obtenidos al aplicar los ítems a una muestra de individuos, pero ambos paradigmas poseen notables diferencias, como se muestra en la Tabla 1 elaborada por Muñiz (2010) y que puede servir como guía para saber cuál es la más adecuada en una situación concreta. Como se aprecia en dicha tabla, para resolver los problemas citados de *invarianza de las mediciones* e *invarianza de las propiedades del test* que no encontraban una buena solución dentro del marco clásico, la TRI utiliza *modelos* más complejos y hace unas *asunciones* más fuertes y restrictivas que las realizadas por la TCT que, por tanto, son mucho más difíciles de cumplir con los datos recogidos. También es distinta la *escala de las puntuaciones* de los test, que es mucho más intuitiva en la TCT. Respecto al cálculo de los valores de los parámetros de *descripción de los ítems*, mientras que la TRI necesita programas de ordenador específicos y una muestra amplia y representativa de personas, matemáticamente la TCT realiza unos cálculos mucho más asequibles basados principalmente en la estadística descriptiva que solo requieren de programas que soporten funciones estadísticas básicas y el *tamaño muestral* necesario puede ser mucho más pequeño que para la TRI.

Aspectos	Teoría Clásica	Teoría de Respuesta a los Ítems
Modelo	Lineal	No Lineal
Asunciones	Débiles (fáciles de cumplir por los datos)	Fuertes (difíciles de cumplir por los datos)
Invarianza de las mediciones	No	Sí
Invarianza de las propiedades del test	No	Sí
Escala de las puntuaciones	Entre cero y la puntuación máxima en el test	Entre - ∞ y + ∞
Énfasis	Test	Ítem
Relación Ítem-Test	Sin especificar	Curva Característica del Ítem
Descripción de los ítems	Índices de Dificultad y de Discriminación	Parámetros a, b, c
Errores de medida	Error típico de medida común para toda la muestra	Función de Información (varía según el nivel de aptitud)
Tamaño Muestral	Puede funcionar bien con muestras entre 200 y 500 sujetos aproximadamente	Se recomiendan más de 500 sujetos, aunque depende del modelo

Tabla 1 – Diferencias entre la TCT y la TRI (Muñiz, 2010).

Así, cuando se utiliza la TRI, aunque el modelo de Rasch requiere relativamente pocos sujetos (Linacre, 1994), para los modelos logísticos de dos y tres parámetros se recomiendan muestras mayores. En caso del modelo 3PL, muchos autores coinciden en recomendar que cada ítem del banco sea administrado a una muestra de por lo menos 500 sujetos (Bunderson, Inouye y Olsen, 1989; Hambleton y Swaminathan, 1985; Renom y Doval, 1999).

Mientras que la TCT considera el test como unidad fundamental, la TRI pone el *énfasis* en el ítem. De acuerdo con Hambleton, Swaminathan et al. (1991a) son varias 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. Igualmente, sus 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 mediante la *función de información*. 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 una misma escala (Ozen y Reise, 1994). A pesar de estas ventajas, los modelos de TRI de ningún modo invalidan el enfoque clásico, si bien constituyen un excelente complemento que en determinadas circunstancias dan solución a problemas mal resueltos en el marco clásico. De hecho, prácticamente todos los test editados en España están desarrollados y analizados dentro de este último marco clásico (Muñiz, 2010).

IV Estandarización de recursos educativos

Los *estándares e-learning* son patrones, modelos y/o reglas relacionados con el campo del aprendizaje electrónico y mayoritariamente aceptados. Su utilización proporciona múltiples ventajas a la comunidad de aprendizaje online puesto que permite (Fallon y Brown, 2003): (1) crear contenidos reutilizables; (2) realizar una migración sencilla de un sistema a nuevas versiones e incluso a una nueva plataforma; (3) comunicarse e intercambiar información con otros sistemas; (4) administrar la información apropiada tanto del recurso como del estudiante; (5) extender los servicios y las capacidades de las plataformas; y (6) asegurar por mayor tiempo la inversión en la infraestructura.

En la actualidad existen varias organizaciones dedicadas a la especificación y establecimiento de estándares *e-learning*. Algunas de ellas – como ADL, AICC e IMS GLC – se centran en la representación de los recursos educativos, también conocidos como *objetos de aprendizaje* (OA). Otras – como DCMI e IEEE LTCS – se centran en la descripción de las características de los OA para clasificar, buscar y recuperarlos de forma más precisa.

El trabajo de todas estas organizaciones ha permitido que actualmente exista en la comunidad e-learning un conjunto de normas técnicas estándares para representar, estructurar y documentar OA, de forma que puedan ser compartidos y reutilizados con mayor facilidad, rentabilizando su construcción (Hernández, 2003).

Este capítulo trata de la estandarización de los OA y comienza con una descripción general de las iniciativas más importantes junto con los logros más destacados de las mismas tanto en la representación de OA como en la representación de sus metadatos (sección IV 1). El resto del capítulo se centra en los estándares de la iniciativa IMS GLC, que son los que usa Hezinet, incidiendo en los que corresponden a los contenidos de evaluación: ítems y test. Con este objetivo, primero se describe el marco de trabajo que utiliza IMS para obtener los estándares (sección IV 2). A continuación, se revisan las especificaciones IMS para el intercambio de contenidos de aprendizaje (sección IV 3) y para el modelado de ítems y test (sección IV 4).

IV 1 Iniciativas de estándares e-learning

En lo referente a la *representación de los OA*, destacan las iniciativas de ADL, AICC e IMS, que se describen a continuación.

ADL (*Advanced Distributed Learning*) son las siglas de la iniciativa de aprendizaje distribuido avanzado (ADL, 2017b) y desarrolla e implementa tecnologías de aprendizaje a través del Departamento de Defensa de los Estados Unidos de América y el gobierno federal. Uno de sus mayores éxitos es el *modelo de referencia de objetos de contenidos compartibles,* más conocido como ***SCORM*** (ADL, 2017a), uno de los más utilizados en el aprendizaje electrónico. La mayoría de las especificaciones de SCORM se han tomado de otras organizaciones como AICC, IMS e IEEE.

AICC (*Aviation Industry Computer-Based Training Committee*) son las siglas del Comité de Entrenamiento Basado en Ordenador de la Industria de la Aviación. Esta iniciativa desarrollaba normativas para sus proveedores de formación basada en ordenador con objeto de garantizar la armonía entre los requisitos de los cursos, y la homogeneización de los resultados obtenidos en los mismos. Estas recomendaciones técnicas, denominadas AGR, abarcan desde la entrega de contenidos hasta los dispositivos periféricos de una formación basada en ordenador. La más seguida de todas ellas es la *AGR 010 dedicada a la interoperabilidad de las plataformas de formación y los cursos*. Sus estándares fueron absorbidos por ADL en diciembre de 2014.

IMS GLC *(Instructional Management Systems Global Learning Consortium)*, generalmente abreviado como IMS (IMS-GLC, 2017) es una organización global sin ánimo de lucro que se centra en la interoperabilidad de sistemas y contenidos de aprendizaje y en la integración de dichas capacidades en la empresa. Los contextos que se benefician de las especificaciones de IMS incluyen entornos específicos de Internet (tales como sistemas de gestión de cursos basados en la Web), así como situaciones de aprendizaje que implican recursos electrónicos off-line (por ejemplo, el acceso del alumno a los recursos de aprendizaje en un CD-ROM). IMS ha aprobado y publicado más de 20 estándares sobre tecnología del aprendizaje, que están disponibles de forma gratuita a través de su página web y se pueden utilizar sin royalties. Entre ellos se encuentran IMS Content Packaging (IMS CP) o IMS Common Cartridge (IMS CC) destinados a la creación de paquetes que posibilitan el intercambio de contenidos de aprendizaje; así como, IMS Question and Test Interoperability (IMS QTI) para el modelado de ítems y test. Además, algunas de sus especificaciones se emplean en estándares de otras iniciativas, por ejemplo, IMS CC se usa en ADL SCORM.

En cuanto a la *representación de metadatos* de los OA, las iniciativas más sobresalientes se deben a DCMI e IEEE LTCS.

DCMI *(Dublin Core Metadata Initiative)* son las siglas de la iniciativa de metadatos Dublin Core (DCMI, 2017), que desarrolla distintos estándares centrados en metadatos. Su mayor logro ha sido establecer un conjunto de 15 elementos – clasificados en contenidos, propiedad intelectual e instanciación – que permiten describir recursos de información y que se ha convertido en uno de los estándares más extendidos en la recuperación de información en la web, ya que proporciona un estándar simple para facilitar la búsqueda, compartición y gestión de información. En este estándar se define el *marco de trabajo de Singapore para perfiles de aplicaciones Dublin Core* (DC-SingaporeFR, 2017), que caracteriza el panorama de los metadatos en términos de niveles de interoperabilidad entre aplicaciones.

IEEE LTSC *(Institute of Electric and Electronic Engineers Learning Technology Standards Committee)* es el comité de estándares en tecnología del aprendizaje del Instituto de Ingenieros Eléctricos y Electrónicos (IEEE, 2017) de Estados Unidos, que es la mayor organización internacional, sin ánimo de lucro, para

fomentar la innovación tecnológica y la excelencia ofreciendo publicaciones, conferencias, actividades educativas y estándares a través de sus asociaciones y que está accesible online. En junio de 2002, el LTSC ha desarrollado uno de los estándares más conocidos y utilizados en el campo del *e*-learning: el estándar de metadatos para los objetos de aprendizaje (*Learning Object Metadata*) o ***LOM*** (IEEE, 2002). En él se define un esquema conceptual de datos en formato XML, que permite describir estos objetos mediante nueve tipos de metainformación: general, ciclo de vida, meta-metadatos, técnica, educacional, derechos de copyright, relación, anotación y clasificación. Y por tanto facilita la búsqueda, evaluación, adquisición y utilización de los objetos educativos. LOM se incluye en las especificaciones de IMS y de ADL. Auspiciado por el comité europeo de normalización, el borrador del estándar LOM se ha traducido al castellano, dando lugar a LOM-ES (Anido y Rodriguez, 2002) y a su perfil de aplicación, certificado por la Asociación Española de Normalización y Certificación en su norma UNE 71361:2010 (AENOR, 2010).

IV 2 Marco de trabajo IMS

El *marco de trabajo* que define IMS en su guía de mejores prácticas (IMS-CP, 2001) para abordar la estandarización de contenidos, se divide en tres partes principales (Figura 4): empaquetado de contenidos (*Content Packaging*), almacén de datos (*Data Store*) y entorno en tiempo de ejecución (*Run Time Environment*). Mediante este marco se define el intercambio de contenidos de aprendizaje entre una herramienta de autor y la herramienta que interpreta la descripción del contenido (el componente LMS de la figura).

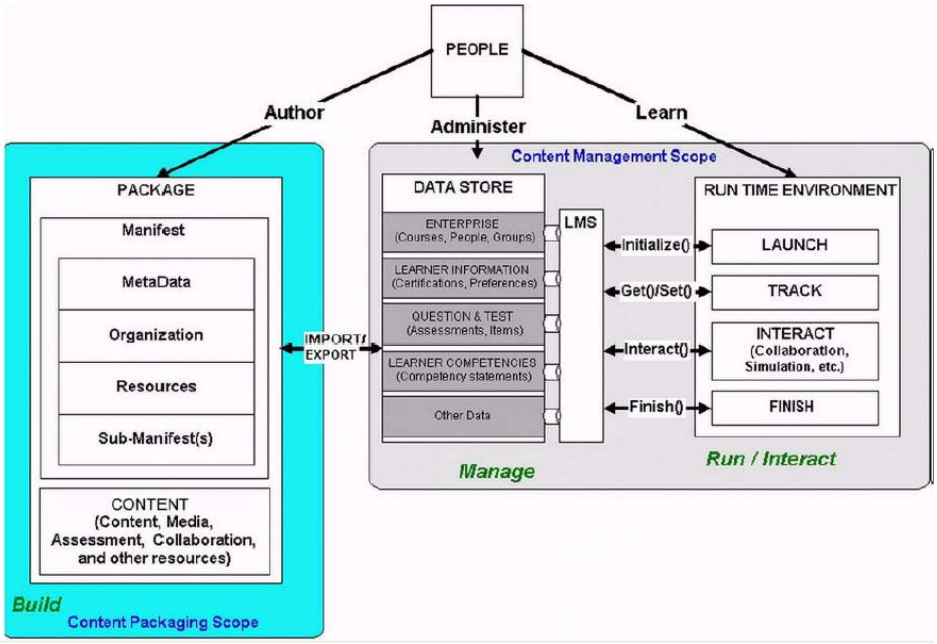


Figura 4 – El marco de contenidos IMS (IMS-CP, 2001).

El *empaquetado de contenidos IMS* representa una unidad de contenido de aprendizaje significativa reutilizable, mediante su encapsulación en un paquete. La estructura de un paquete de contenidos IMS (*PACKAGE*) se compone de dos elementos principales (Figura 4): los contenidos locales a empaquetar (*CONTENT*) y un manifiesto de nivel superior que los describe (*Manifest*). Esta descripción incluye los metadatos (*Metadata*) para el paquete en su conjunto; la estructura para los contenidos (*Organization*); las referencias a todos los recursos reales y elementos que se necesitan (*Resources*); y opcionalmente manifiestos anidados (*Sub-manifest*). Un paquete puede contener varias formas estáticas de organizar los recursos educativos para su presentación. Puede, por ejemplo, describir parte de un curso que puede funcionar por sí mismo fuera del contexto del curso, un curso completo, o una colección de cursos. Los *recursos* descritos son los activos físicos, tales como páginas web, archivos multimedia, archivos de texto, objetos de evaluación (como ítems y test) u otros datos en forma de archivo. También pueden incluir activos que están fuera del paquete, pero disponibles a través de una URL, o colecciones de recursos. IMS trabaja con dos estándares: *IMS Content Packaging* (Smythe y Nielsen, 2007) e *IMS Common Cartridge* (Riley y Mills, 2008), que representan dos métodos de distribución de contenidos reutilizables y que se describen en la siguiente sección.

El *almacén de datos* representa los datos que se utilizan en un entorno de aprendizaje. Entre ellos se encuentran datos de organización empresarial como cursos, grupos o usuarios, las preferencias y certificaciones de los usuarios. Además, permite almacenar la evolución del alumno haciendo referencia a competencias adquiridas o contenidos visitados. También describe los contenidos que se presentan a los usuarios y formas de hacer las presentaciones como, por ejemplo, *IMS Question and Test Interoperability (IMS QTI).*

Los paquetes de contenido sirven como método de transporte de contenidos entre sistemas e-learning. Así, desde el almacén de datos se puede exportar contenido a un paquete de contenidos o importar contenido desde un paquete al almacén de datos. Además, los datos del almacén variarán de acuerdo con las interacciones realizadas por el alumno con el entorno en tiempo de ejecución.

El *entorno en tiempo de ejecución* es el sistema que interactúa con el alumno ofreciéndole la vista de un sistema interactivo de aprendizaje. Cada interacción puede suponer el acceso al almacén de datos para recuperar contenidos, actualizar el perfil del alumno, etc.

El trabajo realizado en cada estándar IMS se plasma en un modelo de datos completo que lo define. Este modelo, denominado *Modelo de Información*, se describe de manera abstracta utilizando UML para facilitar su implementación en diversas herramientas de modelado de datos y lenguajes de programación; y se incluye también una descripción según el estándar XML para el intercambio entre sistemas. Así, cada estándar IMS se define mediante su modelo y, una vez aprobado por la comunidad, se publica online mediante al menos tres documentos: *Information Model* (el Modelo de Información en UML), *XML Binding Specification* (el enlace con el formato XML) y *Best Practices and Implementation Guide* (la guía de implementación y mejores prácticas). Además, en algunos casos están disponibles los esquemas correspondientes a los XML (en formato DTD - Document Type Definition ó XSD - XML Schema Definition) y algunos ejemplos.

IV 3 Estándares IMS para el empaquetado de contenidos

En una estandarización de OA, el empaquetado abarca la descripción, estructura y ubicación de los materiales de aprendizaje online y la definición de algunos tipos de contenido específicos. Las *especificaciones de empaquetado de contenidos* son las que proporcionan los métodos para describir y empaquetar los materiales de aprendizaje haciendo posible su distribución a cualquier sistema compatible con la especificación. En el estándar IMS, tanto la especificación del paquete (IMS CP) como la especificación del cartucho (IMS CC) describen las estructuras de datos que pueden utilizarse para intercambiar datos entre sistemas que desean importar, exportar, agregar y desglosar paquetes de contenido y se apoyan en otros estándares de IMS como IMS QTI para la representación de ítems y test.

La especificación del empaquetado de contenidos IMS CP se centra en la definición de un *paquete lógico* independiente de la plataforma, que representa una unidad de contenido educativo reutilizable. El Modelo de Información correspondiente en UML se basa en el modelo de *manifiesto* (IMS-CP, 2017) y parte de un manifiesto de nivel superior que encapsula metadatos, organizaciones y referencias a recursos (ver Figura 5). Los *metadatos* del manifiesto guardan información acerca del manifiesto en sí, pero también sobre otros dos elementos: organizaciones y recursos. Entre los datos se referencia el esquema que se utiliza (incluyendo la versión empleada) y el metamodelo de datos. Las *organizaciones* del manifiesto representan la estructura organizativa de los recursos empaquetados. Indican una secuencia de estructuras jerárquicas del contenido, la relacionan con los recursos especificados y, posiblemente, con los metadatos asociados. Los *recursos* del manifiesto son las descripciones de los elementos multimedia, archivos de texto, gráficos y otros contenidos educativos que se incluyen en el paquete. Pueden estar organizados en subdirectorios. Los recursos también pueden ser referencias a archivos remotos y metadatos asociados.

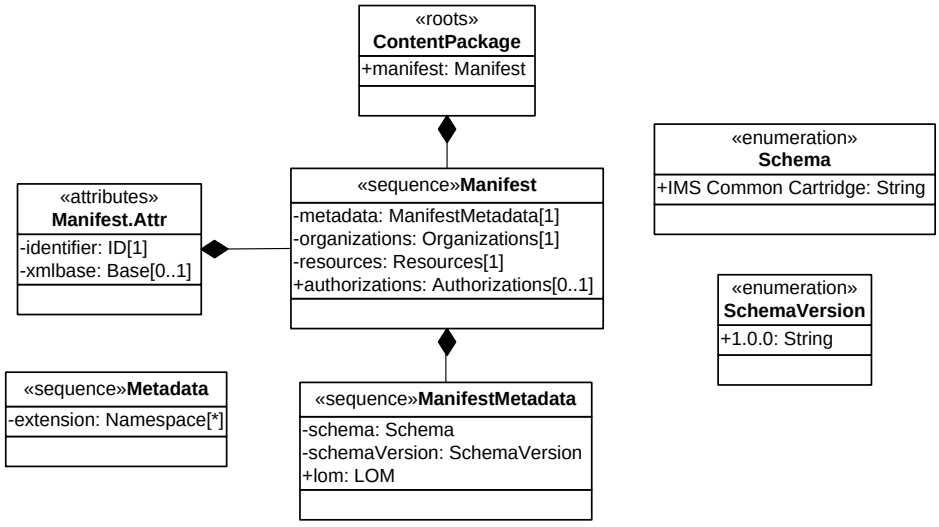


Figura 5 – Modelo para el *manifiesto* de un paquete IMS CC v1.0.0. (IMS-CC, 2017)

Por su parte, la especificación de cartuchos es una evolución del estándar ADL SCORM que aporta mejoras en el ámbito de actividades colaborativas como foros, interconexión con herramientas externas y cuestionarios. Esta especificación consiste en la descripción de ciertas restricciones (denominadas perfiles) a otras especificaciones IMS, como IMS CP e IMS QTI, con objeto de garantizar la máxima interoperabilidad permitiendo solamente el uso de las características básicas de estos estándares (ASPECT, 2017; IMS-CC, 2017). Así, el *perfil del cartucho IMS CC para IMS CP* establece una serie de restricciones que se aplican a esta especificación. En la Figura 5 se detalla el modelo UML para el manifiesto de nivel superior de un cartucho IMS CC versión 1.0.0 tal y como se define en la propia especificación del estándar. Concretamente, un cartucho IMS CC *sólo permite el manifiesto de nivel superior* y para capturar metadatos a nivel de cartucho se utiliza IEEE LOM (clase *MetadataModel*). Además puede contener como máximo una única organización jerárquica. La raíz del paquete (clase *ContentPackage*) se representa mediante el objeto Manifest de nivel superior (clase *Manifest*) y los metadatos asociados a él (clase *ManifestMetadata*). En el cuerpo del manifiesto se indica la organización del contenido (elemento *organizations*) y la ubicación de los recursos necesarios (elemento *resources*).

El paquete lógico IMS abarca el conjunto completo de elementos descritos en el manifiesto de nivel superior que puede incluir a su vez otros submanifiestos. El contenido del paquete se incorpora en un único archivo comprimido para su transporte, que se denomina *paquete de intercambio de ficheros* (PIF), en el que se incluye al propio manifiesto y a todos los componentes locales descritos en él. Así, el PIF es el medio para transportar la información relacionada y estructurada mediante el manifiesto de nivel superior, además de un formato conciso de entrega a través de la web. Para facilitar la creación y modificación de este tipo de paquetes, han aparecido un conjunto de editores que permiten realizar estas operaciones de manera visual e intuitiva. Uno de los más utilizados es RELOAD (http://www.reload.ac.uk/). Un ejemplo de herramienta de código abierto que permite crear recursos y exportarlos en paquetes IMS CP o cartuchos IMS CC es eXeLearning (http://exelearning.net/). Adicionalmente IMS ofrece una herramienta para la validación de paquetes y cartuchos IMS en la web (www.imsglobal.org/developers/alliance/conformancevalidator.cfm).

IV 4 Estándar IMS para el modelado de ítems y test

La especificación IMS QTI permite la representación de *ítems* y de *test*. En concreto, provee un formato de contenido para almacenar e intercambiar ítems y baterías de test (mediante su inclusión en un paquete o cartucho IMS) y sistemas con la habilidad de informar de los resultados de las evaluaciones mediante test de manera consistente; permite crear test interactivos, los cuáles pueden incluir pistas (información para ayudar a los alumnos) y generar diferentes exámenes a partir de una misma plantilla creada previamente (Fernández-Manjón, Moreno-Ger, Sierra y Martínez-Ortiz, 2007; IMS-QTI, 2002, 2017; Smythe et al., 2002).

En IMS QTI la representación de un test se realiza de manera completamente independiente de las preguntas que lo componen, ya que siempre se especifica su funcionamiento a través de las partes o las secciones (Fernández-Manjón et al., 2007).

Para ello, el estándar maneja los conceptos básicos de *test*, *sección* e *ítem.* En la versión 1 del estándar (IMS-QTI, 2002) estos tres conceptos básicos se denominan *Assessment*, *Section* e *Item* respectivamente y, por este motivo, el modelo IMS QTI correspondiente se conoce también con el nombre de modelo ASI. En la segunda versión (IMS-QTI, 2017) se mantiene Section como objeto nuclear y se redefinen los otros dos objetos nucleares, incorporando al estándar los conceptos de test de evaluación (*AssessmentTest*) e ítem de evaluación (*AssessmentItem*).

Tanto el ítem como el test tienen su propio modelo de representación en IMS QTI. Cada uno de estos modelos permite representar todas las características del concepto involucrado, tanto a nivel de estructura como de funcionamiento. El metamodelo ASI para el estándar IMS QTI es el de la Figura 6.

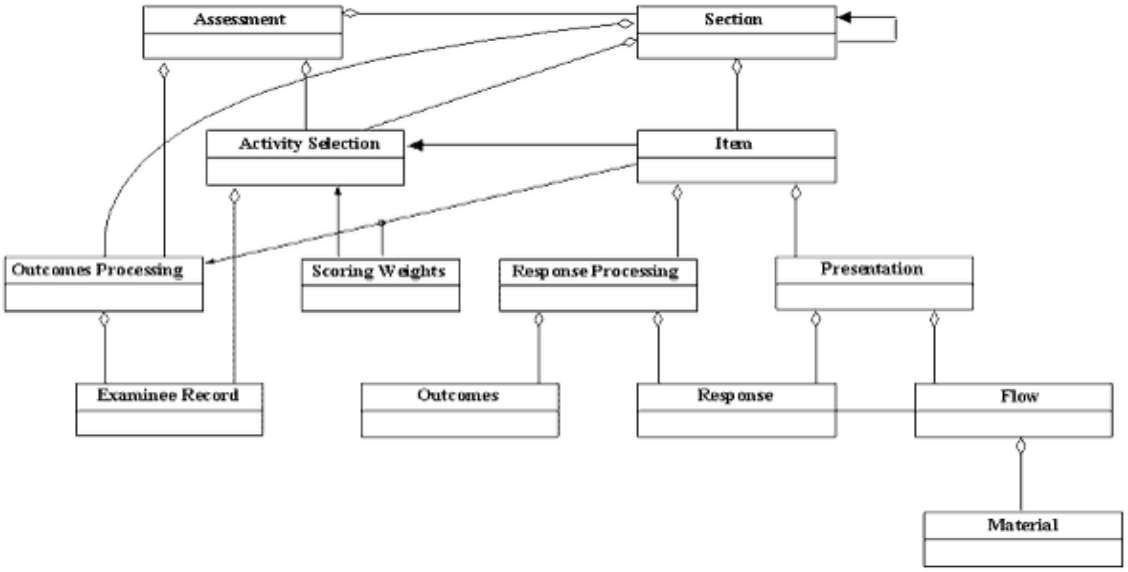


Figura 6 – Metamodelo ASI para el estándar IMS QTI (Smythe et al., 2002)

En cuanto al **modelo de ítem**, entendido como el objeto intercambiable más pequeño dentro de IMS QTI, permite definir cada pregunta del test de manera individual (clase *Item*). El modelo soporta multitud de tipos de preguntas diferentes, incluyendo preguntas simples, de texto y gráficas, pudiendo incluso añadir a las preguntas archivos multimedia. Existen 6 tipos básicos de preguntas: selección múltiple con respuesta única, selección múltiple con respuesta múltiple, verdadero/falso, ensayo, rellenar el hueco y comparación de patrones. Además, IMS QTI permite describir otra serie de elementos asociados a un ítem que aparecen reflejados en la Figura 6 y se detallan a continuación. A nivel de ítem el *procesado de respuestas* (clase *Response Processing*) especifica cuál es el procedimiento de evaluación de las respuestas de usuario. La *presentación* (clase *Presentation*) describe cómo se va a mostrar el ítem. Las *salidas* (clase *Outcomes*) son variables que definen el conjunto de resultados que son evaluados por los objetos de procesado. Estos determinan las métricas de puntuación que son aplicadas a las evaluaciones. Cada instancia de *respuesta* (clase *Response*) contiene las respuestas dadas por el usuario a los *items*. El *flujo* (clase *Flow*) define la estructura de presentación subyacente y la relación entre los diferentes componentes de *material* (clase *Material*) que representan el contenido que se muestra.

De este modo, IMS QTI permite describir la propia pregunta y el resto de elementos asociados a ella. Concretamente, mediante la *presentación* y el *procesado de respuestas* se describe el contenido del ítem (enunciado de la pregunta, posibles respuestas, etc.), cómo se va a mostrar ese contenido en el sistema de entrega, cuál es el procedimiento para puntuar la respuesta del alumno a la pregunta y dónde se almacenará. Opcionalmente, se puede incluir *retroalimentación* asociada a la pregunta (incluyendo pistas y soluciones para el alumno) y especificar *metadatos*, es decir, datos sobre el propio ítem. Para ilustrar este modelo, la Figura 7 muestra la especificación de un ítem, identificado como EUSK1, en formato IMS QTI.

	Especificación del ítem EUSK1 en IMS QTI Versión 1.2
1	<item title=”EUSK1” ident=”EUSK1”>
2	<itemmetadata>
3	<qtimetadata>
4	<qtimetadatafield>
5	<fieldlabel>cc_profile</fieldlabel>
6	<fieldentry>cc.multiple_choice.v0p1</fieldentry>
7	</qtimetadatafield>
8	</qtimetadata>
9	</itemmetadata>
10	<presentation>
11	<material> <mattext texttype="text/html">Zuek zer zarete?</mattext> </material>
12	<response_lid ident="RESPU1" rcardinality="Single"> <render_choice>
13	<response_label ident="A">
14	<material> <mattext texttype="text/html">Gu ikasleak gara.</mattext> </material>
15	</response_label>
16	<response_label ident="B">
17	<material><mattext texttype="text/html">Zuek ikasleak zarete.</mattext> </material>
18	</response_label>
19	<response_label ident="C">
20	<material> <mattext texttype="text/html">Gu ikasleak dira.</mattext> </material>
21	</response_label>
22	<response_label ident="D">
23	<material> <mattext texttype="text/html">Haiek ikasleak dira.</mattext> </material>
24	</response_label>
25	</render_choice> </response_lid>
26	</presentation>
27	<resprocessing>
28	<outcomes>
29	<decvar vartype="Decimal" varname="SCORE" maxvalue="1" minvalue="0"/>
30	</outcomes>
31	<respcondition>
32	<conditionvar> <varequal respident="RESPU1">A</varequal> </conditionvar>
33	<setvar varname="SCORE" action="Set">1</setvar>
34	</respcondition>
35	</resprocessing>
36	</item>

Figura 7 – Ejemplo de ítem en formato IMS QTI

En la zona de metadatos (líneas de 2 a 9), se especifican metadatos sobre la versión de IMS QTI utilizada por el ítem e indican que se trata de un ítem de selección múltiple que sigue el perfil del cartucho IMS CC, concretamente un ítem de selección

múltiple. La presentación del ítem (líneas de 10 a 26) indica el enunciado de la pregunta a responder (línea 11), se especifican 4 opciones de respuesta posibles (etiquetadas como A, B, C y D respectivamente) que se mostrarán una debajo de otra y de las que habrá que seleccionar solamente una, que se almacenará en una variable llamada RESPU1 (líneas de 12 a 24). El procesado de la respuesta (líneas de 27 a 35) indica que la puntuación se guardará en una variable llamada SCORE y que será un número real entre 0 y 1. Cuando la respuesta dada sea la opción A (que es la correcta) esta puntuación será 1, en cualquier otro caso se mantendrá a 0. No se incluye realimentación para el alumno. Concretamente, la especificación de la Figura 7 daría como resultado el ítem EUSK1 presentado en la Figura 8, con las características ya indicadas.

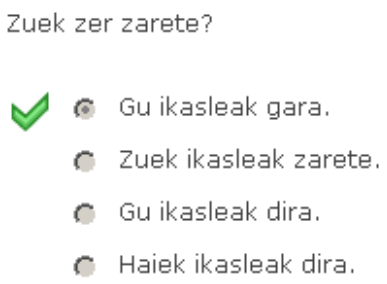


Figura 8 – Ítem de selección múltiple EUSK1 con una sola respuesta correcta.

En cuanto al **modelo de test**, en IMS QTI la especificación de un test (clase *Assessment* de la Figura 6) se realiza siempre a través de la especificación de las secciones. Por tanto, todo test contendrá una sección por lo menos.

Una **sección** representa el concepto nuclear de agrupación y consta de uno o más ítems o secciones (o referencias a ellos). Cada sección (clase *Section* de la Figura 6) tiene sus propias reglas de selección y ordenamiento. En general, los objetos contenidos dentro de una sección tienen alguna relación entre ellos, ya sea en términos de contenido o de entrega del contenido. Dentro de un test, la sección se utiliza para cubrir dos necesidades diferentes: (1) representar agrupaciones tal y como se definen en el propio paradigma educativo, por ejemplo una sección podría ser un subtema; y (2) restringir el alcance de las instrucciones de secuenciación y controlar las formas en que pueden construirse las posibles secuencias diferentes. Así, el objetivo de las secciones dentro del test es doble: por un lado, cuando el alumno realice el test se le presentan cada una de las partes en el orden en el que aparecen en la definición y por otro, cuando se define el test, las distintas secciones se pueden configurar por separado.

IMS QTI permite describir otra serie de elementos asociados a un test o una sección que aparecen reflejados en la Figura 6 y se detallan a continuación. Aparte de la estructuración, un objetivo adicional del estándar es la generación de una única nota para el test que agrupe todas las puntuaciones individuales de las preguntas, ponderándolas con algún factor si fuera necesario. Mediante los pesos de la puntuación (clase *Scoring Weights*) se especifican los pesos asociados a cada una de las respuestas. Para ello, durante la creación del test se puede definir cómo ha de realizarse la agrupación de las puntuaciones individuales. El modelo además incluye una especificación de las posibles formas de navegar a través de las preguntas (v.g. en orden, sin restricciones, etc.), ofrecer retroalimentación, puntuar las diferentes preguntas y establecer restricciones de tiempo (inferiores y superiores) a nivel de test o de

cualquiera de sus componentes (clase *Outcomes Processing*). Por último, la existencia de un registro persistente que guarda los progresos de los alumnos (clase *Examinee Record*) permite indicar cómo se seleccionará la siguiente actividad de acuerdo con el progreso y resultados obtenidos hasta el momento de realizar la selección (clase *Activity Selection*).

Aunque los entresijos son complicados, existen herramientas que permiten la creación, modificación y gestión de cuestiones y test en IMS QTI. Una de ellas es Respondus (http://www.respondus.com/).

En cuanto al empaquetado de ítems y test en IMS, mientras que la especificación IMS CP no impone ninguna restricción al formato IMS QTI a emplear, el *perfil del cartucho IMS CC para IMS QTI* establece que ítems y test deben estar definidos utilizando la primera versión de IMS QTI, y además restringe esa primera versión del estándar. Las restricciones más destacables que impone el cartucho a IMS QTI son: que soporta exclusivamente los 6 tipos básicos de preguntas, que una sección solamente puede contener una lista de ítems – lo que implica que en una sección no se puede hacer referencia a otros ítems – y que la evaluación ha de constar obligatoriamente de una única sección.

V Bancos de ítems calibrados

A partir de la década de los 80, los bancos de ítems han ido sustituyendo cada vez más a los test convencionales (van der Linden, 1986). Atendiendo a su nivel de complejidad, Umar (1997) distingue tres tipos de bancos: un *banco de ítems básico* viene a ser cualquier colección de ítems bien redactados y que son válidos para un determinado fin; un *banco de ítems de nivel de validación empírica tradicional* adjunta las características psicométricas clásicas de cada ítem, tales como los índices de dificultad y discriminación o la distribución de las respuestas; y un *banco de ítems calibrado*, el más complejo de los tres, es aquél que utiliza la TRI para validar los ítems y componer los test de evaluación.

Un *banco de ítems calibrado* (Umar, 1997) es una colección de ítems, que está almacenada de manera estructurada junto con sus características, tanto de contenido como psicométricas, y cuyos parámetros han sido estimados y transformados a una escala común mediante un proceso de equiparación. La información que se guarda en el banco suele incluir, por cada ítem, un identificador único, el enunciado y sus materiales asociados (tablas, fotos, vídeos, sonidos, etcétera), las opciones correctas e incorrectas, una referencia al dominio específico sobre el que evalúa, el número de veces que ha sido administrado y los parámetros del modelo de la TRI al que corresponden. También se suele disponer de índices y tablas que agilicen la búsqueda del ítem más apropiado a cada necesidad e incluso datos de carácter más general como el nombre del autor o la fecha de creación de cada ítem.

Este capítulo se centra en los bancos de ítems calibrados, tanto por expertos como por la TRI, que se utilizan en los sistemas de evaluación y aprendizaje. Concretamente, el capítulo detalla los ámbitos históricos de aplicación de los bancos de ítems calibrados por expertos (sección V 1), para seguir con la utilidad actual de este tipo de bancos en el contexto de los sistemas de aprendizaje y, en este contexto, se relata la construcción de los TAI como una de sus aplicaciones más notables (sección V 2). Para finalizar se describe la calibración de un banco de ítems en el marco de la TRI (sección V 3). Por su parte, el proceso de calibración por expertos, que es en el que se centra esta memoria, se analizará en detalle a lo largo del capítulo VI .

V 1 Ámbitos de aplicación

Respecto al *ámbito de aplicación* de los bancos de ítems calibrados, en este capítulo únicamente se citan algunos de los bancos de ítems de mayor envergadura y de uso real, construidos y calibrados en primera instancia por expertos. Históricamente la primera experiencia documentada surgió durante el periodo 1905-1908, cuando 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 test 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 test 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 educación, en grandes organizaciones médicas, en las Fuerzas Armadas y en grandes compañías de test. En este sentido, donde se ha observado un mayor desarrollo ha sido en el *campo educativo*, por ejemplo, se han empleado en el aula para construir test que permitan a los docentes evaluar el nivel de conocimientos de sus alumnos (Nitko y Hsu, 1984; O'Brien y Hampilos, 1988) o 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*. En la actualidad, los proyectos internacionales de este tipo con más renombre son los 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 International Association for the Evaluation of Educational Achievement): OCDE/PISA (Programme for International Student Assessment, www.oecd.org/pisa/), IEA/TIMSS (Trends in International Mathematics and Science Study, www.iea.nl) e IEA/PIRLS (Progress in Internacional Reading Literacy Study, pirls.bc.edu/). Las pruebas de evaluación del rendimiento académico con dimensión internacional llevadas a cabo por los proyectos PISA, TIMSS Y PIRLS, se realizan de forma periódica (con ciclos que varían de entre los dos y los cinco años), abarcando tanto diferentes áreas de conocimiento como distintos tramos de edad y, en todas ellas, el volumen de la población a la que se aplican (una muestra superior al 75% de la población adecuada por país en cada prueba) es tal que todos estos proyectos precisan necesariamente de bancos de ítems de considerable tamaño. En la preparación de las pruebas colaboran los Ministerios de Educación de los países participantes en el proyecto y existe un consejo representativo de todos esos países, que 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 en grupos de trabajo encargados de preparar y actualizar los bancos de ítems sobre las diferentes áreas de conocimiento a evaluar.

Relacionados con estos proyectos internacionales, va en aumento la difusión de *proyectos nacionales e internacionales* que precisan de bancos de ítems para compilar y administrar sus respectivos test, 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* International iNdicators of Education Systems, *OCDE/PIAAC* Programme for the International Assessment for Adult Competencies, *OCDE/TALIS* Teaching and Learning International Survey, *IEA/TEDS-M* Teacher EDucation Study in Mathematics, *IEA/ICCS* International Civic and Citizenship Study, y *EBAFLS* European Bank of Anchor Items for Foreign Language Skills.

Existen otras muchas evaluaciones de características similares, esta vez para determinar *el nivel lingüístico de los evaluados*, que consisten en la administración de ítems elaborados por expertos a través de pruebas y criterios de corrección parecidos. Por ejemplo, y únicamente de nivel B1 según el Marco Común Europeo de Referencia para las Lenguas del Consejo de Europa o CEFR (Council_of_Europe, 2001) y para alumnos de 6º curso de educación primaria, están el Preliminary English Test (PET) gestionado 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*, cada vez son más las materias que disponen de amplios bancos de ítems para medir las destrezas y habilidades adquiridas por sus alumnos (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 Interno Residente), para los cuales se dispone de extensos bancos de ítems. Por ejemplo, las pruebas de admisión a las universidades de Suecia, las *SweSAT* (Swedish Scholastic Aptitude Test, http://www.edusci.umu.se/english/swesat), se vienen empleando desde 1977 con la TCT como marco teórico subyacente de las pruebas (Stage, 2003) y se celebran dos veces al año. En la actualidad, en estas pruebas participan anualmente cerca de 75000 personas, un equipo de 14 expertos renueva cada dos años los test de las SweSAT, que abarcan las áreas de matemáticas, biología, química y física. La TCT se mantiene como marco teórico, ya que su esquema actual no recomienda el uso de la TRI en su confección, si bien existen ya trabajos para determinar si la TRI permitiría mejorar la calidad de las SweSAT.

Asimismo, y fuera ya del ámbito de la educación reglada y de los bancos de ítems transnacionales, existen *bancos de ítems nacionales* de gran envergadura para realizar *evaluaciones en las oposiciones a gran escala*, como son las oposiciones de distintos niveles para sanidad, administraciones públicas y docentes. Los ítems de dichas pruebas son redactados por expertos en el dominio de 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). Además de los ya citados, se han desarrollado bancos de ítems *con diversos propósitos* en innumerables países del mundo, incluyendo Estados Unidos (Burke, Kaufman y Webb, 1985; Henning, 1986), Australia (Cornish y Wines, 1977; Hill, 1985; Tognolini, 1982), Alemania (Weber, Kuhl y Weibelzahl, 2001), Austria (Kubinger, 1985), Escocia (Pollit, 1985), Holanda (van Thiel y Zwarts, 1986), Inglaterra (Choppin,

1981; Elliot, 1983), y España (Barbero y Navas, 1995; Caro, 1988; Conejo, Guzmán, Millán, Trella, Pérez-De-La-Cruz y Ríos, 2004; Conejo, Guzmán y Pérez, 2008; Pérez, 2000; Trella, Carmona y Conejo, 2005).

V 2 Utilidad en los sistemas de aprendizaje

Desde principios 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. El desarrollo de estos tipos de sistemas está en plena expansión, y el lector interesado puede consultar diversas recopilaciones como las expuestas en Kubes (2007), Weibelzahl (2002) y López-Cuadrado (2010). Las *posibilidades y beneficios* que pueden obtenerse de los bancos de ítems calibrados a la hora de construir test y organizar el dominio de conocimiento de sistemas de aprendizaje (Arruabarrena, Sanz-Santamaría y Gutiérrez, 2007; Barbero, 1996; Olea y Ponsoda, 2003; Wright y Bell, 1984), se resumen en que:

- introducen flexibilidad en la evaluación tanto en el campo psicológico como en el educativo;
- facilitan la comparación de resultados entre evaluados, entre los ítems del banco y entre test;
- permiten reducir el tiempo de administración de los test; pronosticar resultados de la administración de test; 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; y la construcción de test 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 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;
- posibilitan la construcción de *test adaptativos informatizados* (TAI).

La última de ellas es la que ha motivado esta tesis doctoral, esto es, la utilización de bancos de ítems calibrados para la construcción de TAI. Los denominados *test adaptativos* surgen en un contexto en el que es posible estimar los niveles de rasgo de diferentes examinados en la misma escala, incluso aunque no se les haya presentado ningún ítem en común. Según la TRI, ése es precisamente el objetivo de la evaluación: ordenar de algún modo los ítems de más fáciles a más difíciles, y hacer lo mismo con los examinados; esto es, clasificarlos según su nivel de habilidad, empleando para ello la misma escala que con los ítems. Como consecuencia, para realizar evaluaciones de adquisición de conocimiento no es necesario presentar todos los ítems a todos los sujetos. Basta con utilizar los que estén cerca del nivel de habilidad que se está midiendo. Los TAI son la implementación de esta idea en un programa informático que automáticamente presenta los ítems y recoge y evalúa las respuestas. Cualquier programa diseñado para construir TAI debe estar capacitado para predecir, a partir de las respuestas obtenidas hasta el momento, cómo contestaría el evaluado a cualquiera de

los ítems del banco que aún no se le han presentado; elegir conforme a dichas predicciones el ítem más apropiado para ser administrado a continuación; y suministrar tras la finalización del test una valoración numérica que represente la habilidad medida para el examinado (Lord, 1974). Ejemplos de aplicaciones software en las que se ha implementado TAI son el sistema web SIETTE (http://www.siette.org), que es una herramienta desarrollada en la universidad de Málaga para aplicar TAI a través de Internet, y el paquete comercial FastTest (https://www.assess.com/). Para más información sobre las estrategias seguidas en los TAI y cómo se puntúan se puede consultar López-Cuadrado et al. (2010).

En cualquier caso, todo TAI se administra siguiendo un *algoritmo de aplicación*, es decir, respetando unas reglas que definen cuál es el orden en que se le van a presentar los ítems al evaluado (Thissen y Mislevy, 2000). Concretamente, la administración de un test adaptativo comienza una vez se haya seleccionado el primer ítem que se va a suministrar. Esta tarea no es trivial: en primer lugar se deberá disponer de una estimación inicial de la habilidad del evaluado, para posteriormente aplicar un procedimiento que ayude a elegir el ítem que mejor se ajuste a ella. En todo algoritmo de aplicación de un TAI debe establecerse también un método estadístico que resuelva cuál es el siguiente ítem a administrar. Una vez evaluada cada respuesta dada por el examinado, es necesario reajustar los valores estimados de la habilidad del sujeto y de la precisión obtenida. En este punto, el algoritmo deberá seleccionar y mostrar el ítem que mejor se adapte a la nueva estimación de la habilidad. La idea intuitiva que siguen las estrategias adaptativas consiste en que cuando el examinado responde bien a un ítem (o a un conjunto de ítems) el siguiente será un poco más difícil, mientras que como consecuencia de un fallo se suele seleccionar una pregunta más fácil (Hambleton, Zaal y Pieters, 1991b). Es necesario, por último, definir un criterio de parada que determine cuándo ha de finalizar el proceso iterativo de administración de ítems. Por todo lo descrito, los bancos de ítems calibrados son un elemento vital para la construcción del TAI, cuya calidad mejora si se dispone de gran cantidad de ítems adecuados para ser elegidos durante la aplicación del test (Olea y Ponsoda, 2003).

V 3 Calibración de ítems basada en la TRI

Cuando se utiliza como marco teórico la TRI, la *calibración de un banco de ítems* consiste en establecer en una métrica común los valores de los parámetros que definen la curva característica de cada ítem del banco. Aunque la TRI define hasta cuatro parámetros, en la práctica sólo se utilizan los modelos de uno (1PL – dificultad), dos (2PL – dificultad y discriminación) y tres parámetros (3PL – dificultad, discriminación y pseudoacierto) que se describieron en la sección III 2.2. Estos modelos matemáticos se fundamentan en variables (parámetros) latentes, difícilmente observables pero que se pueden estimar. Y en esto consiste precisamente la calibración psicométrica del banco de ítems: se trata de administrar las preguntas a una muestra de sujetos, cuyas habilidades son en principio desconocidas, para obtener estimaciones de los parámetros de cada ítem a partir de las respuestas recopiladas. Para poder asegurar que estos parámetros sólo dependen del ítem y no, por ejemplo, de los sujetos a los que se ha administrado, la muestra utilizada ha de ser lo suficientemente grande y

heterogénea como para que las estimaciones obtenidas sean insesgadas. Así, el primer paso en el proceso de calibración consiste en administrar cada ítem a una muestra de varios cientos de personas. Llevar a cabo una administración de semejantes características obligará probablemente a repartir los ítems entre diversos subtest. Por ello, una calibración de ítems según la TRI se ejecuta por lo general en varios pasos consecutivos (Armendariz, 2014; López-Cuadrado, 2008; Renom y Doval, 1999), que se muestran en la Tabla 2.

Paso	Calibración de ítems basada en la TRI
1	Diseño del experimento a realizar
2	Administración de los ítems
3	Análisis previos (fiabilidad de las administraciones)
4	Estimación estadística de los parámetros de los ítems y de las habilidades de los sujetos
5	Análisis de ajuste al modelo de la TRI (fiabilidad de los ítems)
6	Equiparación de puntuaciones

Tabla 2 – Pasos en la calibración de ítems basada en la TRI.

En el primer paso – *diseño del experimento* – se toman todas las decisiones necesarias para llevar a cabo la calibración, por ejemplo, el número de parámetros a calibrar o los análisis a realizar. En el segundo paso – *administración de los ítems* – se prepara un sistema que distribuye los ítems en distintos subtest, generalmente utilizando algún tipo de *diseño de anclaje* (que consiste en establecer un conjunto de ítems común a todos los subtest), y después se conduce un experimento en el que se administran esos subtest a una gran muestra de sujetos. En el tercer paso – *análisis previos* – se analizan las respuestas recopiladas, para detectar administraciones anómalas (por ejemplo, aquéllas que han sido demasiado rápidas o que presentan la misma respuesta en todos los ítems). Como resultado de esta etapa puede ocurrir que algunas administraciones sean descartadas. En el cuarto paso – *estimación de parámetros* – se estiman estadísticamente los parámetros de los ítems y las habilidades de los sujetos. Para realizar esta estimación estadística existen diferentes técnicas de *estimación conjunta de parámetros y habilidades*, como son las máximo verosímiles y las bayesianas, y uno puede despreocuparse de realizar todos los cálculos a la hora de calibrar un banco de ítems, pues existen paquetes de software que las implementan (López-Cuadrado, Pérez y Armendariz, 2005). En el quinto paso – *análisis de ajuste* – se efectúan estudios de ajuste de los datos consistentes en verificar si las estimaciones recién obtenidas concuerdan con el modelo de la TRI elegido y si se cumplen las restricciones que impone el mismo. La más importante es la comprobación de *unidimensionalidad*, que consiste en verificar que los ítems sirven para medir una única habilidad. Este supuesto puede estudiarse antes de la estimación de parámetros, quedando para después otro tipo de estudios como los de *bondad de ajuste* de los parámetros de los ítems, los de invarianza de los parámetros, o los de simulación del comportamiento del modelo. Como resultado de esta etapa puede ocurrir que algunos ítems sean retirados del banco por no respetar los supuestos de la TRI. Por último, en el sexto paso – *equiparación de*

puntuaciones – se unifican las escalas de los diferentes ítems de anclaje para que todo el banco de ítems (y los test generados a partir de él) utilicen la misma métrica. La equiparación es un proceso estadístico que permite ajustar las puntuaciones de diferentes test, cuyas dificultades probablemente serán desiguales, con el fin de poder compararlas en una escala común.

Aunque no es lo habitual, **la calibración de ítems basada en la TRI también puede recurrir a la TCT** cuando por uno u otro motivo las muestras utilizadas no son lo suficientemente grandes. En este caso, puede resultar conveniente recurrir a los índices clásicos de dificultad (*p*) y discriminación (*r*), que con 150 ó 200 sujetos se pueden estimar de manera estable, como indicadores de los parámetros de los ítems para el modelo de la TRI elegido (Huang, Kalohn, Lin y Spray, 2000). Chang, Hanson y Harris (2000) presentan un procedimiento basado en esta idea que puede servir como primera aproximación en el caso de que no se disponga de muchas respuestas. Por ejemplo, López-Cuadrado y Armendariz (2006) efectuaron un análisis clásico de fiabilidad del ítem tanto para los ítems del conjunto de anclaje como para los ítems de cada uno de los subtest en una calibración psicométrica, y como resultado se obtuvieron los indicadores de fiabilidad correspondientes.

Por último, un aspecto importante a tener en cuenta durante el desarrollo del banco de ítems calibrado es la determinación de su tamaño inicial. Aunque no existen límites inferiores ni superiores concretos para el **tamaño del banco** a calibrar, se pueden seguir algunas recomendaciones. Nunnally (1978) sugiere que, si es la primera vez que se utiliza el banco de ítems, inicialmente éste incluya al menos el doble de los ítems que se espera tener al final, ya que durante el proceso de calibración es habitual que se identifiquen algunos de ellos como erróneos y, en consecuencia, sean retirados. Otros autores recomiendan definir el tamaño mínimo del banco conforme a la longitud de los test que se espere generar a partir de él (Millman y Arter, 1984; Ree, 1977; Stocking, 1994; Way, 1998). Lo ideal, tras el proceso de calibración, sería disponer de una amplia colección de ítems que abarquen todo el espectro de dificultad/habilidad pero, como ocurre con cualquier otro ideal, uno ha de conformarse con la mejor aproximación que esté a su alcance.

VI Proceso de calibración mediante el juicio de expertos

La calibración mediante el juicio de expertos, que estima la dificultad de los ítems a partir de valoraciones subjetivas solicitadas a los 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 literatura que desvele la labor desarrollada por los expertos involucrados en una calibración de ítems. Sirvan como ejemplo las pruebas a gran escala presentadas en el capítulo anterior, que se efectúan con ítems construidos y calibrados, en primera instancia, por expertos en las diferentes áreas de conocimiento a evaluar. Entre ellos, el proyecto PISA es el que más ha transcendido y mejor documentado está, puesto que tanto los análisis como los ítems liberados se recogen en informes nacionales e internacionales, que están dirigidos sobre todo a los administradores y gestores de la educación (Sjøberg, 2004). Aún así, no se ha hallado descrito en documento alguno el proceso concreto llevado a cabo por los comités de expertos para la generación de los ítems calibrados.

Este capítulo propone y detalla un proceso a seguir para calibrar la dificultad de un banco de ítems basándose en las aportaciones subjetivas de expertos. En primer lugar, se abordan los fundamentos a considerar en una calibración de ítems que incorpore la participación de expertos (sección VI 1). En segundo lugar, se define el proceso integral de calibración de ítems con independencia del tipo de participantes y, a continuación, se particulariza este proceso para el caso de una estimación de la dificultad mediante el juicio de expertos (sección VI 2). Se finaliza detallando el diseño del proceso a través de la pormenorización y discusión de las tareas que deben ser planificadas y ejecutadas a lo largo del mismo (sección VI 3).

VI 1 Fundamentos sobre la calibración de ítems con expertos

La bibliografía específica para calibración vía expertos es prácticamente inexistente, aunque existen salvedades como el banco de ítems del test de inteligencia de Binet-Simon a principios del siglo pasado y el estudio PISA en la actualidad. Del proceso seguido, se sabe que los expertos de PISA calibran inicialmente los ítems y también se encargan de actualizar los bancos de ítems, garantizando que los materiales renovados tengan cualidades de medición sólidas y que los instrumentos pongan énfasis en la autenticidad y validez educativa (Mullis, Martin, Kennedy y Foy, 2007; OECD, 2013). Posteriormente, una vez administradas las pruebas que contienen los ítems, en

los estudios PISA, y en la mayoría de las evaluaciones internacionales, se utiliza la metodología de la TRI para comparar resultados de alumnos a nivel nacional e internacional y confrontar resultados de evaluaciones anteriores, gracias a que no todos los ítems empleados en un mismo ciclo de evaluación son descartados (MEC, 2007; Mullis et al., 2007) ya que se conservan para próximas evaluaciones. De las evaluaciones PISA 2009 y 2012 (OECD, 2012, 2014), que se describen en sendos informes técnicos que ofrecen detalles sobre la implementación y la tecnología utilizada, se puede acceder a información relativa a los procesos de construcción y calibración de ítems así como de equiparación de puntuaciones. Por ejemplo, en la evaluación PISA de 2009 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 (OECD, 2012). Sin embargo, el proceso concreto desarrollado por los expertos para construir los ítems calibrados no se ha hallado documentado, encontrándose únicamente alguna referencia superficial sobre el mismo.

Por otro lado, hace ya un tiempo que el testimonio de los expertos es permisible como fuente de conocimiento científico en aquellos ámbitos donde no es posible definir leyes científicas explícitas (Helmer y Rescher, 1959). Además, dicho testimonio es indispensable cuando el empleo de métodos alternativos a los expertos para manejar problemas de cierta enjundia puede involucrar procesos prohibitivos en la práctica, en términos de costes y tiempo, de recolección y procesamiento de la información. Estas justificaciones son válidas hoy día para muchas aplicaciones cuando no se dispone de la información precisa, es muy costoso conseguirla o la evaluación requiere de datos subjetivos en sus principales parámetros. Por otro lado, 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 generan nuevo conocimiento científico gracias al uso de clases, taxonomías, ontologías, reglas fuzzy, heurísticos y redes bayesianas. Todas estas aplicaciones, junto con sus publicaciones, corroboran 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 este motivo, *el problema de la participación de expertos en un proceso se reduce a cómo obtener su testimonio y a cómo combinar el testimonio de varios expertos en una declaración única*.

Es precisamente en la bibliografía referente a sistemas informáticos en los que han participado individuos en calidad de expertos o especialistas en el dominio de conocimiento de la aplicación, donde se han buscado y han aparecido 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 están relacionados con la planificación del proceso, con los propios expertos y con la validez de los resultados.

En primer lugar, la importancia de la **planificación del proceso** es incuestionable, no solo en los procesos de calibración de ítems sino también en todos aquellos procesos que consumen muchos recursos. Esta planificación ha de ser lo suficientemente exhaustiva, correcta y eficiente. La captación, gestión y control de los

recursos humanos, así 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 los debidos a Scriven (1991), Tessmer (1993), y Worthen, Sanders y Fitzpatrick (1997) y en los ámbitos de la evaluación y los experimentos empíricos controlados se puede citar entre otros a Wohlin, Runeson, Höst, Ohlsson, Regnell y Wesslén (2000).

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 (OECD, 2013) o bien a nivel teórico las *guías* propuestas en Worthen 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 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 realista, que considere el número de expertos participantes y el formato cualitativo o cuantitativo de los datos (Arruabarrena y Pérez, 2010).

En cuanto a los aspectos a tener en cuenta **sobre los propios expertos**, algunas publicaciones sobre desarrollos donde han colaborado expertos consideran sumamente importante dar indicaciones sobre la forma en la que éstos deben aportar su conocimiento y el número de expertos participantes, incluyendo la posibilidad de abandono y en qué medida se daría. En el primer aspecto, algunas *técnicas de evaluación o herramientas para obtener información* de una serie de personas son los paneles de expertos, las entrevistas, las encuestas o cuestionarios, el benchmarking, los test, los grupos de enfoque, los mapas conceptuales, y el método Delphi. 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 en Harvey (1998) y Scriven (1991) o también en libros de interacción persona computador, como por ejemplo en Dix, Finlay, Abowd y Beale (1998), Nielsen y Mack (1994) o 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, Shneiderman (1998) puntualiza que las revisiones y participaciones formales de expertos han demostrado ser eficaces, como por ejemplo, en los estudios descritos por Jeffries, Miller, Wharton y Uyeda (1991) y Karat, Campbell y Fiegel (1992). El gran inconveniente de estos métodos es la necesidad de tener que disponer, bien en plantilla o bien como asesores externos, de expertos o profesionales del área. Además, surge la duda de *cuántos expertos es necesario 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 et al. (1998) recomienda 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 y Cochran, 1970), donde se empleaba el método Delphi, mostraron que 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 entre 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 superior a 13 expertos, aunque considerando la relación coste-beneficio, el número óptimo de expertos a involucrar es de 7. Fijada esta cantidad óptima de expertos, otro factor importante a considerar es que *hay que prever el porcentaje de los que iniciarán el proceso pero que no lo concluirán*. Este abandono del experto puede ocurrir principalmente en calibraciones en las que participa de forma no remunerada (Arruabarrena y Pérez, 2010). Si este es el caso, resulta obligado comenzar el desarrollo con un número significativamente mayor de expertos voluntarios que el inicialmente establecido como adecuado.

Respecto a la **validez de los resultados**, 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, con objeto de mitigar posteriores amenazas. 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 a llevar a cabo. En particular, dos son los problemas a menudo presentes en las críticas al uso de herramientas para obtener información de expertos: 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 diagramas PERT. No obstante, y aunque no es habitual medir la bondad de las aportaciones de los expertos, los más escépticos mantienen que los juicios de expertos pueden ser mero reflejo de apreciaciones personales condicionadas (Worthen et al., 1997). Asímismo, la *selección de la muestra de expertos* es otro aspecto que va de la mano de la valía de éstos. 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, donde se afirma que "la selección de un grupo tan diverso como sea posible minimiza el sesgo debido a la selección no aleatoria de los expertos" (Lang, 1995). Cuando existen distintos grupos de expertos consultados, para verificar la validez de los resultados se pueden emplear índices Kappa, intervalos de confianza y porcentuales entre los valores de las estimaciones generadas por cada grupo.

En la valía de los resultados, otra cuestión clave es la fiabilidad de los procedimientos de medición empleados, por lo que 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 dicho con menos rigor, la tendencia de un observador a dar consistentemente valores mayores o menores que otro – y por otra parte, la *concordancia entre observadores (confiabilidad)* o coincidencia de los observadores en su medición. Para realizar el contraste de confiabilidad, el método más frecuentemente empleado es el índice Kappa-Fleiss (Fleiss y Cohen, 1973). El análisis de fiabilidad de la calibración basada en expertos debiera incluir el estudio de ambos aspectos. En este sentido, Arruabarrena (2010) lo incorpora en la familia de criterios empleada durante el análisis de datos, tal y como se vio en la sección II 3.

VI 2 Proceso de calibración de ítems vía expertos

Este apartado comienza presentando un proceso integral de calibración que es común a los métodos basado en la TRI y basado en expertos. Este proceso es el resultado del estudio realizado sobre los procesos diseñados durante las calibraciones llevadas a cabo sobre el banco de ítems de Hezinet, que se han mencionado en la introducción de esta memoria.

La Tabla 3 presenta los diversos agentes involucrados en dicho proceso general de calibración (Arruabarrena, 2010). En el proceso intervienen una serie de individuos que se pueden clasificar en dos tipos: Los *participantes activos* (responsable, supervisor, coordinador, ejecutor y otros colaboradores) – que se encargan de llevar a cabo el proceso y de poner en marcha las distintas tareas a realizar durante la calibración – y los *participantes pasivos* (revisores y sujetos administrados), con los que contactarán los sujetos activos para recabar información.

Agentes involucrados
Supervisor general y responsable del proceso integral
Coordinador y ejecutor principal
Otros colaboradores activos: Informático, transcriptor de cuestionarios en formato papel, ayudantes para la construcción de las pruebas de campo por centros de trabajo o bien en laboratorios
Revisores: del material pedagógico, de los cuestionarios y de la aplicación informática administradora de cuestionarios electrónicos
Sujetos administrados: expertos o bien individuos anónimos

Tabla 3 – Agentes involucrados en la calibración de ítems.

En este contexto se denominará *responsable de la calibración* o simplemente *responsable* a un único agente que adopta todos los roles activos de la Tabla 3, con lo que la labor de este agente es doble: por un lado debe tomar todas las decisiones de

diseño del experimento y por otro debe supervisar, coordinar y controlar el desarrollo del proceso una vez comenzada su ejecución.

En cuanto a las tareas que conforman el proceso, si bien este trabajo se centra en la calibración de ítems, la *preparación del banco de ítems* es una tarea previa al proceso de calibración y es idéntica para cualquier tipo de calibración. Consiste en crear los ítems a calibrar y asegurarse de que todos los ítems del banco son correctos desde el punto de vista tanto de contenido como de formato. A continuación, se procede a la fase de *Administración de ítems* en la que se detalla la conducción y la recogida de los datos del experimento. Aunque se pueden emplear multitud de técnicas para recabar información de los sujetos pasivos (Arruabarrena, Pérez, Gutiérrez, López-Cuadrado y Vadillo, 2002; Mark y Greer, 1993) las aportaciones individuales en este contexto se obtendrán mediante administración de formularios (test y/o cuestionarios) a través de pruebas de campo. Así, en esta fase se planifica, prepara y ejecuta una prueba de campo a través de la cual se intenta recabar un volumen de datos establecido. Por último, en la fase de *Análisis de datos y calibración* se filtran los datos recogidos, se estiman los parámetros de los ítems y se analiza la fiabilidad de los resultados. Esta fase parte de los datos obtenidos y puede comenzar una vez que se cuenta con un cierto volumen de datos, no siendo necesario esperar a que concluya la fase de administración. Cada una de estas dos fases debe ser planificada previamente, puesto que el responsable de la calibración deberá tomar una serie de *decisiones de diseño* antes de su ejecución. Por tanto, si se tiene en cuenta esta tarea de diseño, el proceso integral de calibración consta de tres tareas principales (Tabla 4): diseñar el experimento a realizar, ejecutar la administración y ejecutar la fase de análisis y calibración. Estas tres tareas constarán a su vez de otras subtareas que ya dependen del tipo de calibración a ejecutar.

Tarea	Calibración de ítems
1	Diseñar el experimento a realizar
2	Ejecutar la administración
3	Ejecutar la fase de análisis y calibración

Tabla 4 – Tareas principales del proceso integral de calibración.

Las dos primeras decisiones a tomar en el proceso son: qué conjunto de ítems se va a calibrar y qué tipo de calibración – expertos o psicométrica – se va a acometer. Ambas determinan el resto de decisiones a tomar, ya que estas decisiones se refieren a las subtareas concretas de las dos fases del proceso y éstas dependen del tipo de calibración elegido. Las decisiones implicadas en el *Diseño del experimento a realizar* se encuentran clasificadas en la Tabla 5 según su orden temporal dentro del proceso.

Tarea	1. Diseñar el experimento a realizar
1.1	Especificar ítems
1.2	Especificar experimento
1.3	Especificar análisis, estudios y cálculos para calibrar

Tabla 5 – Tareas en el diseño del experimento a realizar.

Las dos primeras especificaciones de la Tabla 5 deben realizarse antes de comenzar la ejecución de la fase de administración. Concretamente, en la primera tarea se debe *especificar el conjunto de ítems* a calibrar y en la segunda se toman el resto de decisiones de planificación *del experimento*. Respecto a la última especificación de la Tabla 5, son decisiones previas a la estimación de los parámetros del ítem a partir de la muestra recogida. Los *análisis, estudios y cálculos para calibrar* describen todos los procedimientos que tratan de atestiguar la fiabilidad de ítems y de los sujetos administrados una vez recogidos todos los datos de las administraciones, para incluirlos o no en el cálculo de los parámetros. En esta última especificación también se indican los detalles de cálculo de los distintos parámetros a calibrar y los criterios que garantizan la fiabilidad de ese cálculo.

El procedimiento integral de calibración de ítems descrito se puede particularizar para el **caso de una calibración vía expertos**, una vez preparado un banco de ítems con una cantidad suficiente de ítems, especificados los ítems a calibrar y el tipo de calibración como basada en juicios de expertos. En este caso se considera solamente un parámetro escalar: la *dificultad* de los ítems, que es el único parámetro fácilmente medible por los expertos de los tres que utilizan los modelos prácticos de la TRI. En cuanto a los participantes, en una calibración de este tipo intervienen como mínimo, un sujeto activo o *responsable* de la calibración que es quien desea obtener los ítems calibrados y los sujetos pasivos o *expertos* a los que se les suministrarán los cuestionarios. Opcionalmente, se pueden utilizar *revisores*.

La fase de *administración* en una calibración de ítems vía expertos consiste en la realización de las pruebas de campo para obtener las valoraciones de la dificultad de los ítems a través de la administración de cuestionarios a los expertos. En este subproceso se recaba la información entrevistando expertos mediante encuestas plasmadas en cuestionarios y consta de dos tareas consecutivas (Tabla 6): la preparación de un sistema para la administración de cuestionarios y la conducción del experimento.

Tarea	2. Ejecución de la administración
2.1	Preparar el sistema de administración
2.2	Conducir el experimento

Tabla 6 – Tarea en la ejecución de la fase de administración vía expertos.

La *preparación del sistema de administración* consiste en crear un sistema para administrar el banco de ítems a los expertos, donde se materialicen todas las especificaciones previas sobre los ítems a calibrar y las decisiones iniciales tomadas sobre el experimento correspondiente a esa calibración. Por ejemplo, se materializa el tipo de comunicación elegida (basada en correo – postal o electrónico – o en un artefacto informático que genere y albergue los cuestionarios en formato electrónico con la estructura planteada al planificarlos) junto con los plazos y métodos de envío, para a continuación *conducir el experimento* localizando a los expertos y distribuyéndoles los cuestionarios para su cumplimentación. Así, el objetivo de la ejecución de la prueba de campo es distribuir los cuestionarios diseñados a los expertos captados, recopilarlos una vez completados y almacenar los resultados para su posterior análisis.

La fase de *análisis y calibración* se centra en la obtención de los valores de dificultad de los ítems a partir de las aportaciones de los expertos obtenidas mediante la

ejecución de la fase de administración. Consta de tres tareas básicas (Tabla 7): el análisis de datos, la calibración en dificultad y los estudios de fiabilidad.

Tarea	3. Ejecución de la fase de análisis y calibración
3.1	Analizar datos
3.2	Calibrar en dificultad
3.3	Estudiar fiabilidad

Tabla 7 – Tareas en la ejecución de la fase de análisis y calibración vía expertos.

El *análisis de datos* consiste en realizar algunas actividades de análisis sobre la muestra recogida apartando datos irrelevantes, erróneos o anómalos que puedan malograr los resultados de la calibración de los ítems, con objeto de garantizar la validez del experimento. Una vez depurada la muestra se procede con la *calibración* propiamente dicha, que consiste en estimar el valor del parámetro dificultad correspondiente a cada ítem que ha superado los análisis de datos. Para ello, se aplica el procedimiento preestablecido de estimación de la dificultad a cada ítem no descartado, es decir, se emplean los valores subjetivos de la variable otorgados por los expertos a ese ítem y que se hallan en la muestra final depurada. Este cálculo es posible, a pesar de que se pueda disponer igualmente, por cada par experto-ítem, de otra estimación (por ejemplo de la destreza evaluada por el ítem), siempre que ambas variables sean independientes y sus respectivos parámetros puedan estimarse separadamente (Pérez, 1999; Wonnacott y Wonnacott, 1991). Los *estudios de fiabilidad* tienen lugar una vez finalizada la recogida de datos y consisten en realizar análisis que avalen los procedimientos de medición empleados y los resultados obtenidos. Es posible que estos estudios de fiabilidad no constituyan una actividad aparte, sino que la fiabilidad de los resultados de la calibración se encuentre implícita en los análisis de datos ya realizados. Así, el objetivo de esta fase de análisis y calibración es realizar los análisis, estudios y cálculos oportunos a la muestra de datos recogidos a los expertos, aplicando los filtros correspondientes en el orden decidido, obteniendo como resultados una muestra final depurada, en la que las administraciones e ítems descartados se hallan etiquetados junto con los descriptores de los criterios que originaron su retirada del banco, y el banco de ítems calibrado en dificultad a partir de esta muestra depurada.

VI 3 Diseño del proceso de calibración

La tarea de *diseño del proceso* consiste en que el responsable especifique todas las decisiones necesarias para llevar a cabo una calibración vía expertos. Así, además de las decisiones sobre el experimento en sí, se han de tomar otra serie de decisiones que tienen que ver con la planificación del desarrollo de las otras dos fases a ejecutar, tanto las subtareas que las compondrán y su orden de ejecución como los resultados a obtener en cada fase del proceso. En esta memoria se aplica este supuesto, con lo que las tareas de ejecución simplemente llevan a cabo las actividades previamente planificadas mediante el diseño.

Este apartado describe y discute todas las decisiones involucradas en el proceso. La sección VI 3.1 se centra en la especificación de los ítems y del experimento (tareas 1.1 y 1.2 de la Tabla 5), mientras que la sección VI 3.2 discute la especificación de los análisis, estudios, cálculos y estimaciones a realizar (tarea 1.3 de la Tabla 5).

VI 3.1 Especificación de ítems y experimento

La especificación de ítems y experimento tiene como objetivo **diseñar la fase de administración**, lo que implica decidir los pormenores de diseño de las dos tareas sucesivas de las que consta esta fase: la preparación del sistema de administración y la conducción del experimento.

Así, durante el diseño de esta fase se decide qué sistema de administración se va a crear, qué tipo de comunicación tendrá lugar, cómo serán los cuestionarios a administrar, cómo se distribuirán esos cuestionarios entre los expertos participantes y qué información aportada por los expertos se almacenará para su posterior análisis. Además, dependiendo del sistema de administración de cuestionarios que se utilice, puede ser necesaria la especificación de algún método que permita recoger la muestra de datos cumplimentados por los expertos, obtenida durante la conducción del experimento, desde ese sistema de administración al sistema que realizará la fase de análisis y calibración.

Por otra parte, es *recomendable iniciar la depuración de la muestra antes de dar por finalizadas las pruebas de campo* ya que los análisis utilizados en la depuración de los datos pueden reducir el volumen de la muestra por debajo del umbral establecido para realizar una calibración fiable. Este particular puede desencadenar una *tarea de compensación que garantice el mínimo de aportaciones por ítem*, que generalmente consiste en dilatar la fase de administración de cuestionarios recogiendo más datos hasta alcanzar el umbral de valoraciones fijado (Arruabarrena, 2010). Una alternativa a esta recogida de datos adicionales cuando existen ítems no descartados con pocas valoraciones válidas, puede ser aplicar a los datos que se tienen otros filtros menos exigentes que conserven más aportaciones.

Una vez especificado el conjunto de ítems a calibrar, se continúa con la especificación del experimento, donde se establecen una serie de decisiones previas a la conducción de la prueba de campo que definirán el marco de funcionamiento del sistema de administración de los cuestionarios a los expertos. Todas ellas se discuten en los siguientes párrafos y se han clasificado en tres tipos, siguiendo a Arruabarrena y Pérez (2010): especificaciones iniciales de planificación, especificaciones sobre los expertos participantes en el experimento y especificaciones sobre el diseño y revisión de los cuestionarios.

En las **especificaciones iniciales de planificación** se indican las decisiones sobre el proceso que sirven como base para la toma de las demás decisiones. Se deben identificar los objetivos del proceso de calibración (parámetros de calibración a estimar y volumen de datos a recoger); estudiar el entorno donde se va a desarrollar; definir los entregables a generar y escoger las herramientas y técnicas a utilizar en el proceso.

En cuanto a los *entregables*, se debe decidir qué material se va a administrar, cómo se entregará este material a los expertos dispuestos a colaborar, cuáles serán las fechas de entrega y los plazos de media para completar los formularios y recogerlos. Por

otro lado, el estudio del *entorno* permite determinar el perfil de los expertos que se pueda tener disponibles, su preparación y su tipo de participación. Esta última cuestión es particularmente importante a la hora de estimar cuántos expertos se necesitan, puesto que si su participación es voluntaria y no remunerada es más probable que no se involucren con el rigor esperado y será aconsejable suponer una elevada tasa de abandono y porcentaje de descarte.

También se deben decidir las *herramientas* de administración de cuestionarios y las *técnicas* de cálculo que se utilizarán. Si se prevé que la alfabetización informática de los expertos puede incidir negativamente en la participación, se debe sopesar el recopilar la información entrevistando a expertos mediante encuestas plasmadas en papel en vez de utilizar cuestionarios electrónicos (Dix et al., 1998; Tessmer, 1993). Respecto a las herramientas informáticas a utilizar para apoyar el proceso global de calibración, a día de hoy existen multitud de paquetes informáticos, tanto para la construcción y administración de ítems y cuestionarios como para tareas concretas del análisis psicométrico y de la calibración de un banco de ítems. Así, para la fase de administración de ítems se dispone de software que abarca desde programas sencillos que realizan tareas concretas de esta fase hasta aplicaciones más complejas capaces de implementarla en su totalidad, como son algunas plataformas educativas web. Por otro lado, como software de apoyo para la fase de análisis estadístico y calibración se puede emplear cualquier paquete estadístico del mercado que sea capaz de realizar los análisis personalizados que se definan en el proceso.

En las **especificaciones sobre los expertos** se incluyen ciertas decisiones sobre el conjunto de expertos tales como la estrategia de selección, identificación y formación. También se debe decidir qué cuestionarios se administrarán a cada uno de los expertos que participen. Para más información se puede consultar la referencia (Arruabarrena y Pérez, 2010).

En las **especificaciones sobre el diseño y revisión de los cuestionarios** la mayor parte de las ocasiones este diseño consiste en fraccionar el banco de ítems en varios grupos. En algunos casos, como el descrito en la sección V 3, se suele repetir un pequeño número de ítems (denominados de anclaje) en todos los grupos con el objeto de facilitar la posterior equiparación de las estimaciones (Kolen y Brennan, 1995). Además, si la participación de los expertos es voluntaria y no remunerada, se recomienda que los cuestionarios se diseñen para poder ser respondidos en un corto espacio de tiempo y que la *duración total del cuestionario a rellenar por un experto no exceda de un tiempo límite*. Esto puede implicar que haya que repartir el banco de ítems en varios cuestionarios y calcular el número adecuado de ítems para ser valorados dentro de ese margen de tiempo. Dependiendo de la cantidad y el perfil de los expertos que se prevea tener disponibles para contestar a los cuestionarios existen distintas estrategias de *distribución de los ítems* en esos cuestionarios. Por ejemplo, puede plantearse un reparto inicial para facilitar esta labor (Armendariz, 2014).

Además de repartir todos los ítems del banco en distintos cuestionarios, para finalizar el diseño de los mismos el responsable debe tomar *otra serie de decisiones antes de su administración* que tienen que ver con los datos que se van a solicitar al experto para cada ítem del banco y a la estructura de los cuestionarios a distribuir. En cuanto a los *datos a recopilar sobre los ítems*, puesto que se utiliza el nivel de dificultad del ítem para calibrar el banco suministrado, éste será en principio el único dato obligatorio a recoger. Por ejemplo, se puede decidir si se solicitará también la respuesta

correcta del ítem para utilizarla como elemento de control o si se permitirá dejar el nivel de dificultad sin estimar. En cuanto a la estructura de los cuestionarios, por ejemplo se puede decidir si se incluirán o no directrices claras de cumplimentación y ejemplos ilustrativos. La inclusión de unas instrucciones de cumplimentación en las que se indique claramente cuál es la escala de valoración a la hora de estimar la dificultad con objeto de que todos los expertos apliquen el mismo criterio, evitará la necesidad de una posterior equiparación de puntuaciones (Arruabarrena, 2010).

Por último, para asegurar la idoneidad de la extensión de los cuestionarios y localizar posibles deficiencias y carencias en los mismos, se puede realizar una *administración a revisores mediante una prueba piloto*, consistente en presentar los cuestionarios diseñados a revisores individualmente para que los comprueben. Como alternativa a los revisores, durante el diseño de los cuestionarios también *puede plantearse la realización de algunos cálculos matemáticos* que alerten cuando la confección de los mismos incumpla ciertas restricciones antes de su distribución. Por ejemplo, pudiera ocurrir que algún ítem no se hubiera incluido en ningún cuestionario.

VI 3.2 Especificación de análisis, estudios y cálculos para calibración

La especificación de análisis, estudios y cálculos para calibración tiene como objetivo **diseñar la fase de análisis y calibración**, lo que implica decidir los pormenores de diseño de las tareas de las que consta esta fase. En este diseño, los análisis de datos pueden incorporar también los estudios de fiabilidad, con lo que las actividades que conforman la fase de análisis y calibración se podrían reducir a dos tareas sucesivas: el análisis de datos y de fiabilidad y la propia calibración.

Concretamente, mediante esta especificación, se deciden y detallan los análisis de fiabilidad que tratan de atestiguar que ítems, expertos y resultados son aceptables junto a los procedimientos de cálculo de parámetros que se centran en el parámetro dificultad del ítem. Estas decisones se dividen en dos: la especificación de análisis y estudios y la especificación del cálculo del parámetro dificultad.

Mediante la **especificación de análisis y estudios** se establecen varios filtros y procedimientos para depurar los datos recogidos de los expertos y mantener aquéllos que permitan obtener el valor de la dificultad de cada ítem de forma fiable, lo que incluye la realización de pruebas para verificar su funcionamiento y orden de aplicación.

Así, las primeras decisiones que tiene que tomar el responsable de la calibración antes de calcular la dificultad final de cada ítem están relacionadas con el descarte de administraciones (entendidas como contribuciones de expertos) y la eliminación de ítems. Para ello, se establecen *tres niveles de filtrado* para descartar aportaciones anómalas o incorrectas y para descartar tanto ítems inválidos (junto con todas sus aportaciones), como expertos (junto con todas sus aportaciones) que no hayan participado en el experimento con el rigor esperado (Arruabarrena y Armendariz, 2008). Estos tres niveles implican la especificación de otros tantos tipos de análisis: de aportación, de ítem y de administración.

Un *análisis de aportación* utiliza una serie de criterios para analizar la validez de cada estimación de nivel dada por cada experto a cada ítem propuesto. El objeto de este

análisis es retirar aquellas entradas inservibles para los cálculos a realizar. Por ejemplo, el criterio C.ex-1 presentado en la sección II 3.

Un *análisis de ítem* analiza la fiabilidad de un ítem particular utilizando las respuestas de los expertos con el objetivo de decidir si dicho ítem posee la suficiente calidad para permanecer en el banco. Por ejemplo, el criterio C.it-1 presentado en la sección II 3.

Un *análisis de administración* analiza para cada cuestionario las aportaciones del experto que lo cumplimentó. El objetivo de este tipo de análisis es detectar y rechazar aquellas administraciones de expertos que perjudican el desarrollo correcto de la calibración. Por ejemplo, el criterio C.ex-2 presentado en la sección II 3. Los análisis de administraciones se hacen especialmente necesarios cuando se trabaja con administraciones no supervisadas. Cuando la administración es supervisada, es el propio supervisor el que asegura ciertas condiciones de cada una de las aplicaciones de los cuestionarios. Si bien, eso no quita que se puedan hacer también este tipo de análisis, aunque sea relajando alguno de los términos.

El responsable es quien decide si va a llevar a cabo estos análisis y tendrá que especificar los criterios que se van a utilizar en cada uno de ellos para depurar la muestra de datos recogidos. Estos criterios podrán ser comunes a ambos tipos de calibración (vía expertos y psicométrica) o ser criterios específicos de las calibraciones vía expertos.

En el primer caso, Armendariz (2014) ha definido varios criterios generales de validación de administraciones en el contexto de las calibraciones psicométricas, algunos de los cuales pueden ser también aplicables a las calibraciones vía expertos. Concretamente, estos criterios son los análisis de *administraciones incompletas* (administraciones que no se llegan a terminar), los análisis de *patrones de respuesta anómalos* (administraciones con respuestas al azar, selección sistemática de una opción de cada ítem, u omisión sistemática de respuestas), los análisis de *tiempos totales extremos* (administraciones que se han realizado en mucho tiempo o por el contrario en muy poco tiempo y no alcanzan un cierto umbral de aciertos), los análisis de *pausas intermedias* (administraciones en las que se ha invertido mucho tiempo en responder al menos un ítem), los análisis de *todo aciertos o todo fallos* (administraciones que presentan bien todas las respuestas correctas o todas falladas) y los análisis de *administraciones rápidas* (administraciones que, estando entre las más rápidas, no superan el número de aciertos medio que ha tenido una administración en toda la muestra). De todos estos criterios, solamente uno sería directamente aplicable para validar administraciones a expertos tal cual está definido: el análisis de administraciones incompletas, en el que no se llega a terminar el cuestionario, esto es, se deja sin valorar un porcentaje de los ítems del cuestionario. Respecto a los demás análisis, su uso en una calibración vía expertos bien no tiene sentido o bien implicaría su adaptación a este tipo de calibración. Por ejemplo, varios de estos criterios se basan en los tiempos mínimos y en que una respuesta instantánea, tan rápida que se pueda suponer que al individuo no le ha dado tiempo material para siquiera leer el enunciado, puede indicar un acierto por casualidad en una administración psicométrica (Mostow, Tobin y Cuneo, 2002). Del mismo modo, en un análisis de administraciones a expertos se podrían considerar este tipo de tiempos mínimos ya que, respuestas demasiado rápidas a los ítems, indicarían una falta de rigor del experto al rellenar el cuestionario.

Por otra parte, también en el contexto de las calibraciones estadísticas, cuando se lleva a cabo un estudio de los porcentajes de omisiones en las respuestas al ítem y se obtienen valores pequeños, las omisiones en las administraciones se pueden considerar como respuestas erróneas sin que ello repercuta negativamente durante la estimación de parámetros (Olea y Ponsoda, 2003). Del mismo modo en una calibración vía expertos con este supuesto, el estudio de las omisiones en la respuesta correcta al ítem cuando es obligatorio, se podría englobar en el análisis de respuestas incorrectas al ítem dadas por el experto.

En cuanto a la especificación de los criterios propios de una calibración vía expertos, se pueden considerar *dos tipos de filtros* siguiendo las recomendaciones de depuración sobre marcos imperfectos (Frechtling y Sharp, 1997; Otero y Dolado, 2007; Pérez, 1999): criterios de aceptación o rechazo de las aportaciones de expertos y criterios de aceptación o rechazo de un ítem. El primer tipo sirve para estudiar la *fiabilidad de los expertos* y es el filtro que salvaguarda el resultado final de aquellas contribuciones que no se hayan realizado con el rigor esperado por parte de los sujetos administrados. Por su parte, el segundo tipo realiza un estudio de *fiabilidad de los ítems* detectando ítems de poca calidad. Es importante que la aplicación de los filtros *no produzca sesgo en la muestra* (Arruabarrena y López-Cuadrado, 2006; Otero y Dolado, 2007), de manera que los criterios empleados deberán limitarse a retirar ítems defectuosos, aportaciones incorrectas o juicios inservibles, salvaguardando la validez de las aportaciones de los expertos. Para *concretar los filtros* a emplear en una calibración basada en expertos el responsable puede inspirarse en métodos estadísticos que permitan evaluar el parámetro a calibrar. Habitualmente, esta inspiración parte del estudio de los indicadores y estadísticos utilizados en el análisis clásico de la TCT, que es especialmente útil en los casos en los que la revisión de ítems no sea posible antes de administrar los test (Muñiz, 2000). Se pueden encontrar ejemplos de criterios y filtros en (Arruabarrena y Armendariz, 2008) y en (Arruabarrena, 2010).

Por último, si el orden de aplicación de los criterios definidos puede alterar la muestra depurada resultante, *se debe especificar su orden de aplicación y concretar los que tienen que plasmarse de forma combinada y/o cíclica*. Con este objetivo, se pueden realizar análisis adicionales sobre la incidencia de aplicar en órdenes alternativos las familias de criterios enunciados y a partir de los resultados obtenidos, depurar la muestra eliminando las aportaciones inaceptables (Otero, 2003; Scanlan, 1989). 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 la referencia (Arruabarrena y Armendariz, 2008).

Finalmente, se pueden establecer más análisis independientes para estudiar la **fiabilidad de procedimientos y resultados**. En principio, se pueden utilizar cualquiera de los métodos mencionados en la sección VI 1. También, para el cometido de fiabilidad de los resultados, si se ha establecido un mínimo de aportaciones a recopilar por ítem, se pueden especificar criterios para asegurar que el volumen de la muestra depurada alcanza el umbral de datos establecido para garantizar una calibración fiable (Armendariz, 2014; Arruabarrena, 2010).

Con respecto a la **especificación del cálculo del parámetro dificultad** a partir de las aportaciones de los expertos, existen dos tipos de alternativas para determinar el nivel de dificultad de cada uno de los ítems. Para empezar, se puede buscar el consenso entre los propios expertos tras dar sus estimaciones iniciales *volviéndolos a reunir*

utilizando, por ejemplo, alguna de las técnicas vistas en la sección VI 1. La otra alternativa, menos costosa, es aquélla que evita la necesidad de volver a consultar a los expertos. Estos últimos procedimientos siempre son estadísticos y su objetivo es establecer el valor más probable entre los pronósticos más consensuados (Arruabarrena, 2010). Una aproximación de este tipo sería utilizar *estadísticos que generen valores discretos* como por ejemplo, la *moda* de las aportaciones de los niveles de dificultad por ítem (Palacios, Pérez, Callejón y Herrerías, 1999). Sin embargo, si existen muchos ítems cuyos pronósticos tengan frecuencias de dificultad repetidas, no es adecuado emplear únicamente la moda como método estadístico puesto que no permite determinar diferencias entre ítems del mismo nivel. Otra posibilidad sería usar *estadísticos que generen dificultades con valores no discretos* como por ejemplo, la *media* para obtener valores continuos. Sin embargo, las valoraciones extremas pueden alterar la estimación, por lo que también se desaconseja emplear solamente la media como estadístico para determinar la dificultad. Por último, se pueden aplicar *estadísticos ad-hoc* como por ejemplo, el estadístico *M.dif* presentado en la sección II 3, que permite que el nivel de dificultad del ítem sea determinado por los pronósticos más frecuentes y no sea distorsionado por juicios extremos.

Parte Tercera: Estado Del Arte

La Parte Tercera se compone de 2 capítulos, en los que se presenta el **estado del arte** en cuanto al software disponible para la administración de test y calibración de ítems, enfocado hacia los que han sido relevantes para el trabajo de investigación recogido en esta tesis.

El capítulo **VII - Software para administración** trata sobre las aplicaciones Web actuales que soportan estándares e-learning centrándose en Moodle, la plataforma utilizada para la construcción del sistema CALLIE-EXPERT.

El capítulo **VIII - Software para análisis de datos y calibración** complementa al anterior y enumera otras herramientas software destacadas de cara a automatizar el proceso de calibración mediante procedimentos estadísticos.

VII Software para administración

Existen en el mercado múltiples herramientas software que permiten automatizar la fase de administración de cuestionarios con distintos grados de cobertura para cada una de las tareas que la conforman. Independientemente de su complejidad, unas herramientas utilizan sus *propios formatos de representación* de ítems y test, como MALTED – Multimedia Authoring for Language Tutors and Educational Development (http://malted.cnice.mec.es), Exam Software (http://www.exam-software.com/), Aritest Profesores (http://www.aritest.com/), HotPotatoes (http://web.uvic.ca/hrd/hotpot/), TestGIP Notas (http://testgip-notas.software.informer.com/) y Form Pilot (http://www.colorpilot.com/); mientras que otras soportan *estándares para la representación*, como Canvas (https://canvas.instructure.com/), Adobe Authorware y Adobe Dreamweaver CC (http://www.adobe.com/es/), Kedros-LMS (http://www.satec.es/) y Toolbook (http://tb.sumtotalsystems.com/). En el contexto de esta tesis doctoral la fase de administración de ítems se lleva a cabo mediante una plataforma LMS que soporta estándares e-learning de interoperabilidad, y que se integra en la herramienta implementada, lo que permite que CALLIE disponga de multitud de capacidades extra durante esta fase del proceso de calibración del banco de ítems. Así, este capítulo revisa exclusivamente el estado del arte de este tipo de plataformas.

Una *plataforma e-learning* es una aplicación web que integra un conjunto de herramientas para la enseñanza-aprendizaje en línea, permitiendo un aprendizaje no presencial (*e-learning*). Es posible también un enfoque de enseñanza mixta (*b-learning*), donde se combina la enseñanza por Internet con experiencias en la clase (PLS-RAMBOLL, 2004). Estas plataformas ofrecen la posibilidad de proporcionar conocimiento en cualquier momento y en cualquier lugar donde haya una conexión a Internet, y su uso ha transformado una parte de los espacios de enseñanza tradicionales en entornos virtuales de aprendizaje o EVA (López Alonso y Matesanz del Barrio, 2009). Actualmente el software más utilizado para crear y dar soporte a múltiples EVA son los *sistemas de gestión del aprendizaje* (*LMS*, de Learning Management System). Un LMS generalmente no incluye posibilidades de creación de contenidos educativos, sino que se centra en gestionar contenidos creados por fuentes diferentes. La creación de estos EVA, normalmente, se realiza utilizando una plantilla que personaliza un conjunto de herramientas que el diseñador considera necesarias para llevar a cabo el proceso de aprendizaje (Clarenc, 2012; Horton y Horton, 2003; López Alonso y Matesanz del Barrio, 2009; Marcelo, 2006) como: (1) herramientas de administración del EVA, (2) herramientas de comunicación, (3) herramientas de gestión de grupos, (4) herramientas de evaluación y (5) herramientas de gestión de contenidos. Dichas herramientas y/o servicios están disponibles en los LMS, y pueden ser añadidos fácilmente a un curso. Todas ellas se describen brevemente a continuación.

Las *herramientas de administración del EVA* permiten realizar las operaciones de creación, borrado y modificación de usuarios del sistema, definición de roles, asignación de tutores, así como llevar un control y seguimiento de los accesos al

sistema. También permiten la creación, borrado y modificación de los EVA que se soportan.

Las *herramientas de comunicación* permiten la comunicación y colaboración entre los diferentes roles que hay en entorno de aprendizaje, como profesores, tutores o estudiantes. Por ello a veces se les denomina herramientas colaborativas. Hay herramientas de comunicación síncrona – si se establece una comunicación en vivo o en tiempo real entre los participantes – o asíncrona – si no es así. Algunos ejemplos de herramientas síncronas son el chat, la pizarra electrónica, el audio y las videoconferencias; y como herramientas asíncronas se pueden citar el correo electrónico, los foros y las wikis.

Las *herramientas de gestión de grupos* permiten realizar las operaciones de alta, modificación o borrado de grupos de alumnos y la creación de escenarios virtuales para el trabajo cooperativo de los miembros de un grupo. Estos escenarios de grupo pueden incluir espacios para el intercambio de archivos, herramientas para la publicación de los contenidos, y foros o chats privados para los miembros de cada grupo.

Las *herramientas de evaluación* permiten la creación, edición y realización de pruebas de evaluación. Las pruebas pueden ser de autoevaluación, de tal forma que el estudiante pueda afianzar o corregir posibles problemas que se le han presentado en el aprendizaje. También pueden permitir al profesor obtener un informe de las respuestas que han dado los estudiantes.

Las *herramientas de gestión de contenidos* también conocidas como *LCMS* (Learning Content Management System) permiten (Fallon y Brown, 2003): generar descripciones únicas para cada objeto de aprendizaje; descubrir (buscar y localizar) el objeto de aprendizaje requerido; proporcionar jerarquías para el almacenamiento y organización de los objetos de aprendizaje; y facilitar la creación de cursos. Así, el LCMS crea los cursos y sus contenidos en forma de módulos que se pueden personalizar, manejar y usar en diferentes ocasiones. Además, proporciona una gestión de contenidos orientada al e-learning integrando generalmente estándares de producción de contenidos educativos reutilizables (como los comentados en el capítulo IV). Un LCMS puede estar integrado en el propio LMS, o los dos pueden comunicarse entre sí.

Toda plataforma e-learning debe estar instalada en un servidor y necesita un software de tipo *servidor Web* que procese la aplicación desde el lado del servidor. Por otro lado, necesita un *servicio de autentificación* ya que, para realizar cualquier acción (participar en cursos, crear cursos, administrar el servidor, etc.), es necesario acreditarse como *usuario* de la plataforma en el servidor. Por último, necesita un *gestor de base de datos* que guarde y gestione en una base de datos toda la información relevante sobre usuarios, recursos, parámetros de configuración, etc. La infraestructura del servidor de instalación puede estar en un servidor del proveedor de contenidos o en la nube y la funcionalidad ofrecerse como un servicio online.

Una plataforma e-learning puede ser creada con fines lucrativos por una compañía comercial o puede seguir la filosofía del software libre. En este último caso no suele haber una empresa comercial detrás, sino que generalmente se crea un proyecto online, con el mismo nombre que el producto, en el que colaboran distintos individuos (desinteresadamente la mayoría). En este tipo de proyectos, el software no se libera al obtenerse un producto final, sino más bien cada vez que se hace una mejora o ampliación estable significativa, lo que lleva a una gran cantidad de versiones

intermedias y a una evolución en las características del producto cada poco tiempo. Dicho software se puede distribuir mediante varios tipos de licencias que indican el grado de libertad que se dará al usuario del software, siendo la más permisiva la licencia GNU/GPL (General Public License), que significa que el código se puede usar, distribuir y modificar sin coste alguno. Así, las plataformas de código abierto, son plataformas gratuitas distribuidas mediante licencia GNU/GPL o compatible, lo que permite a cualquier persona montar y modificar la propia plataforma incluyendo la posibilidad de añadirle nuevas funcionalidades. Además, a diferencia de las plataformas de pago, las de código abierto están extensamente documentadas, tanto a nivel de uso como de desarrollo (código). Si se desea ampliar los conocimientos sobre los distintos tipos y plataformas se puede consultar la referencia (Clarenc, Castro, López de Lenz, Moreno y Tosco, 2013).

En la actualidad existen infinidad de LMS, cada uno de ellos con sus propias herramientas y funcionalidades. Como no es posible encontrarlas a todas en una única plataforma, es importante que en el momento de optar por uno de estos sistemas de gestión de aprendizaje se evalúe qué indicadores son más importantes para su institución, como también con qué presupuesto y recursos (humanos, administrativos y tecnológicos) se cuenta, para tomar una decisión acertada que se ajuste a las necesidades presentes y futuras (Clarenc, 2012, 2013).

En la sección VII 1 se hace una revisión del estado del arte de los LMS más populares que soportan gestión de contenidos de evaluación y que utilizan algún estándar e-learning para garantizar la interoperabilidad de sus contenidos. En consecuencia, todas las plataformas que se citarán permiten al menos ensamblar, empaquetar y redistribuir ítems y cuestionarios mediante algún estándar e-learning y llevar a cabo cursos online. A continuación, en la sección VII 2 se hace una descripción más detallada del LMS Moodle, debido a que esta plataforma ha sido la utilizada en el diseño y construcción del sistema de administración de cuestionarios a los expertos en la herramienta de ayuda presentada en esta tesis.

En este caso, los indicadores a considerar para cada plataforma revisada son: sus posibilidades de ampliación desarrollando funcionalidades propias, la cobertura de estándares de interoperabilidad y el tipo de herramientas que tiene implementadas. Como se deberá disponer de un servidor web propio para CALLIE-EXPERT, es irrelevante que la plataforma ofrezca servicios de hosting en la nube o no. En esta compilación, para cada compañía/proyecto que desarrolla software LMS se indican las siguientes dimensiones: (1) sus características (nombre, lugar, país, año de comienzo) y a qué entorno van dirigidos sus productos (educación o empresa), (2) la descripción de los productos que ofrece, indicando si son productos comerciales (de pago) o de código abierto, (3) las herramientas de que dispone: LCMS, de administración del EVA, de comunicación, de evaluación y/o de gestión de grupos; y (4) los estándares e-learning de interoperabilidad soportados.

VII 1 Revisión de plataformas LCMS con estándares e-learning

En los siguientes párrafos se detallan las distintas dimensiones para las compañías/proyectos responsables de las plataformas LMS más relevantes en la actualidad. Se citan siguiendo un orden alfabético: Atutor, Blackboard, Catedr@, Claroline, Chamilo, Desire2Learn Inc., Docebo Srl, Dokeos, eCollege, Ilias, Moodle, Saba Software y Sakai.

ATutor (http://atutor.ca/) es un proyecto de software libre que comenzó en 2002, coordinado por el ATRC (Adaptive Technology Resource Centre) de la universidad de Toronto (Canadá) y su software se dirige principalmente a instituciones educativas. Ofrece el producto *Atutor* bajo licencia de código abierto. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, posee herramientas de evaluaciones en línea, así como herramientas de autoría y de colaboración. En cuanto a la cobertura de estándares e-learning, el contenido creado en otros sistemas conforme a los estándares IMS o SCORM se puede importar al sistema ATutor, y viceversa. Los estándares e-learning de interoperabilidad IMS que soporta son: IMS CP 1.1.2+, IMS QTI 1.2 y 2.1 e IMS CC 1.0.

Blackboard (http://www.blackboard.com/) es una compañía comercial estadounidense fundada en 1997, con sede central en Washington D.C. y su software se dirige principalmente a instituciones educativas. Ofrece varios productos bajo licencia comercial: Blackboard Learn, Blackboard Collaborate, Blackboard Mobile, Blackboard Connect y Blackboard Analytics lo que permite el acceso desde los dispositivos más populares y la colaboración mediante herramientas de aprendizaje social. *Blackboard Learn* es la plataforma LMS comercial que se puede combinar con Blackboard Collaborate y Blackboard Mobile, para crear comunidades de colaboración e integrarlas. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, posee herramientas colaborativas, de evaluación y de seguimiento y gestión de aprendizaje incluida la gestión de grupos. En cuanto a la cobertura de estándares e-learning, les da soporte mediante un módulo denominado The Open Standards Content Player Building Block que permite importar contenidos conformes a SCORM e IMS en un curso.

Catedr@ (http://www.catedra.edu.co/) es una empresa colombiana creada en el año 1992, con sede en la ciudad de Bogotá y su software se dirige a escuelas, universidades y empresas. Ofrece varios productos bajo licencia comercial: (1) Open School que es LMS y LCMS, un sistema de registro y control académico, biblioteca electrónica, sistema de evaluación por logros y competencias; (2) Open University que es una suite compuesta por un sistema de Registro y Control y un LCMS orientado a las instituciones de educación superior; (3) y Open Company que es una suite orientada al uso de empresas con las funcionalidades de un LMS y un sistema de gestión de recursos humanos con evaluación por desempeños. Su producto comercial LCMS se denomina *catedr@lcms*. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, posee herramientas colaborativas y no hay datos sobre otro tipo de herramientas. En cuanto a la cobertura de estándares e-learning, constituye un sistema de enseñanza y aprendizaje con estándares SCORM.

Chamilo (https://campus.chamilo.org/) es un proyecto de software libre que comenzó en 2010, como una bifurcación del proyecto Dokeos, coordinado por la asociación Chamilo, y su software se dirige a educación y empresas sobre todo latinoamericanas. Ofrece varios productos bajo licencia de código abierto: por un lado ofrece *Chamilo LMS* que tiene dos versiones: Chamilo LMS y Chamilo LCMS Connect; por otro lado, proporciona un campus de e-learning de acceso público y gratuito. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, permite la creación de ejercicios y la generación de informes. En cuanto a la cobertura de estándares e-learning, presenta compatibilidad conforme a SCORM y es posible importar cursos que sigan los estándares SCORM, AICC e IMS.

Claroline (http://www.claroline.net/) es un proyecto de software libre que comenzó en 2000, coordinado por la universidad católica de Lovaina (Bélgica) y su software se centra en el ámbito de la educación. Ofrece el producto Claroline bajo licencia de código abierto. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, posee múltiples herramientas colaborativas, de gestión de grupos y de evaluación. Concretamente dispone de una serie de herramientas para: configurar las características de cualquier curso; publicar documentos en cualquier formato como word, pdf, html, vídeo, etc.; administrar foros de discusión tanto públicos como privados; administrar listas de enlaces; crear grupos de estudiantes; confeccionar ejercicios y test; estructurar una agenda con tareas y plazos; hacer anuncios, vía correo electrónico por ejemplo; gestionar los envíos de los estudiantes: documentos, tareas, trabajos, etc. y crear y guardar chats. En cuanto a la cobertura de estándares e-learning, sigue las especificaciones de SCORM e IMS QTI v2.

Desire2Learn Inc. - D2L (http://www.brightspace.com/) es una empresa canadiense creada en 1999, con sede en Ontario y su software se dirige al entorno empresarial. Ofrece varios productos bajo licencia comercial, con la denominación *Brightspace* (desde 2012, antes *Desire2Learn Learning Suite*) que combina un LMS, un LCMS, un Repositorio de Objetos de Aprendizaje y otras herramientas en línea. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, posee herramientas de evaluación y no hay datos sobre otro tipo de herramientas. En cuanto a la cobertura de estándares e-learning, es compatible con el estándar IMS.

Docebo Srl (http://www.docebo.com/) es una compañía italiana fundada en 2005, con sede en la ciudad de Nápoles, y su software se dirige a PYMEs, corporativos y a instituciones de educación superior, ofreciendo un apoyo a los diversos modelos educativos. Ofrece el producto *Docebo* en dos tipos de plataformas: Docebo Cloud, la cual opera desde una infraestructura en la "nube", y Docebo Premium, que puede ser instalado en los sistemas de otras compañías con una licencia abierta, o bien entregado como un servicio. Además dispone de una versión para teléfonos móviles inteligentes. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, posee herramientas colaborativas. En cuanto a la cobertura de estándares e-learning, es compatible con los estándares y las especificaciones SCORM. Trabaja con software libre.

Dokeos (http://dokeos.com/) es un proyecto de software libre que comenzó en 2004, y su software se dirige a empresas. Ofrece el producto Dokeos bajo licencia de código abierto. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, posee herramientas colaborativas y de evaluación pues incluye: distribución de contenidos, calendario, proceso de entrenamiento, chat en texto,

audio y video, administración de pruebas de evaluación y almacenamiento de registros. En cuanto a la cobertura de estándares e-learning, soporta la importación de archivos en SCORM 1.2.

eCollege (http://www.ecollege.com/) es una compañía estadounidense fundada en 1996, con sede en Denver, actualmente es parte de la compañía Pearson y su software se dirige al entorno universitario. Ofrece varios productos, bajo licencia comercial, dentro de una única solución denominada *Pearson eCollege LearningStudio* que incluye entre otros: un LMS que consiste en Entorno de Aprendizaje, Administrador de resultados de Aprendizaje, Administrador de reportes Ejecutivos, y Class Live Pro; un CMS que consiste en Administrador de Contenidos y Herramienta de Autoría de Cursos; y un Diseño instruccional que consiste en una serie de servicios orientados a diseñar instruccional y pedagógicamente cursos en línea para ser llevados a cabo mediante el LMS y alojados en el CMS. También ofrece una solución gratuita en la nube: *OpenClass*. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, no hay datos sobre otro tipo de herramientas. En cuanto a la cobertura de estándares e-learning, da soporte a la interoperabilidad de contenidos mediante los estándares IMS.

Ilias (www.ilias.de) es un proyecto de software libre que comenzó en 1999, tuvo su primera versión disponible en la red hacia septiembre de 2000, actualmente está coordinado por un equipo de la universidad de Colonia en Alemania y su software se dirige tanto a las empresas como a la educación. Ofrece el producto *Ilias* bajo licencia de código abierto. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, posee múltiples herramientas integradas que permiten una creación eficiente de cursos y material, ofrece aprendizaje y trabajo cooperativo en la plataforma sin herramientas adicionales, múltiples formas de entregar los contenidos de aprendizaje, un entorno integrado para la creación de test y evaluaciones, formas estándar de comunicación como chats, foros y correo electrónico y varios métodos de autentificar a los usuarios de la plataforma. En cuanto a la cobertura de estándares e-learning, soporta los estándares SCORM, AICC, metadatos LOM e IMS QTI.

Moodle (http://moodle.org/) es un proyecto de software libre que comenzó en 2002 en la universidad tecnológica de Curtin (Australia), después de que Martin Dougiamas creara la primera versión de un LMS basado en la idea de que el estudiante construye sus conocimientos a partir del aprendizaje colaborativo y su software se dirige al entorno educativo. Ofrece el producto *Moodle* bajo licencia de código abierto. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, posee múltiples herramientas colaborativas, de gestión de grupos y de evaluación. En cuanto a la cobertura de estándares e-learning, permite importar paquetes SCORM y soporta diversos estándares de IMS, posibilitando, por ejemplo, la importación de cursos en el formato estándar IMS CC mediante el sistema de backup de Moodle.

Saba Software (http://www.saba.com/) es una compañía estadounidense creada en 1997, con sede central en Redwood Shores (California) y su software se dirige a la gestión y formación de recursos humanos en empresas. Ofrece varios productos bajo licencia comercial: Learning, Performance, Planning, Collaboration, Succession y Recruiting. El primero de ellos, *Saba Learning* contiene el software específico para LCMS dividido en tres componentes: Learning Content, Content Management y Publishing Tools. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, posee múltiples herramientas de desempeño, planificación,

colaboración, sucesión de tareas y reclutamiento. En cuanto a la cobertura de estándares e-learning, da soporte completo a la interoperabilidad de contenidos mediante los estándares AICC y SCORM.

Sakai (http://www.sakaiproject.org/) es una fundación sin ánimo de lucro a la que pertenecen más de 100 universidades y cuyo objetivo es crear un entorno de colaboración y aprendizaje (CLE) para la educación superior, que pueda competir con sus equivalentes comerciales como Blackboard y que mejore otras iniciativas de código abierto como Moodle; su software se dirige al entorno universitario. Su proyecto asociado ofrece el producto *Sakai CLE* que es un sistema de gestión de cursos basado en Java distribuido bajo licencia ECL (Educational Community License), que es una licencia GNU/GPL especializada para la comunidad educativa. En cuanto a las herramientas disponibles, además del LCMS y las de administración del EVA, posee múltiples herramientas de comunicación, lector de noticias RSS, distribución de material docente, realización de exámenes y gestión de trabajos. En cuanto a la cobertura de estándares e-learning, cumple los estándares IMS CC 1.0 y 1.1 y permite la creación de contenidos en IMS QTI.

La Tabla 8 muestra la síntesis de las 13 compañías/proyectos estudiados.

Compañía/Proyecto	Orientación	Productos	Estándares soportados	Herramientas	Código abierto
Atutor	Educación	Atutor	SCORM, IMS CP 1.1.2+, IMS QTI 1.2 y 2.1, IMS CC 1.0	LCMS y adm, colaborativas, de evaluación	Sí
Blackboard	Educación Empresas	Blackboard Learn	SCORM, IMS	todas	No
Catedr@	Educación Empresas	catedra@lcms	SCORM	LCMS y adm, de evaluación	No
Chamilo (Asociación)	Educación Empresas	Chamilo LMS	SCORM, AICC, IMS	LCMS y adm, de evaluación	Sí
Claroline	Educación	Claroline	SCORM, IMS	todas	Sí
Desire2Learn Inc.	Empresas	BrightSpace	IMS	LCMS y adm, de evaluación	No
Docebo Srl	Educación superior Empresas	Docebo Cloud, Docebo Premium	SCORM	LCMS y adm, colaborativas	Sí
Dokeos	Empresas	Dokeos	SCORM 1.2	LCMS y adm, colaborativas, de evaluación	Sí
eCollege	Educación superior	Pearson eCollege Learning Studio, Pearson eCollege OpenClass	IMS	LCMS y adm	No
Ilias	Educación Empresas	Ilias	SCORM, AICC, metadatos LOM, IMS QTI	todas	Sí
Moodle	Educación	Moodle	SCORM, IMS, IMS CC, IMS CP	todas	Sí

Compañía/Proyecto	Orientación	Productos	Estándares soportados	Herramientas	Código abierto
Saba Software	Empresas	Saba Learning	AICC, SCORM	LCMS y adm, colaborativas, de evaluación	No
Sakai (Fundación)	Educación superior	Sakai CLE	IMS CC 1.0 y 1.1, IMS QTI	LCMS y adm, colaborativas, de evaluación	Sí

Tabla 8 – Síntesis de compañías/proyectos que soportan estándares e-learning.

En cuanto a la orientación, la mayoría de las compañías estudiadas se dirigen al ámbito educativo y solamente 3 se dirigen exclusivamente al ámbito empresarial. Existen dos que no soportan estándares SCORM y otras cuatro que no soportan el estándar IMS (dos de ellas dirigidas a empresas). Las más completas a nivel de herramientas son Blackboard, Claroline, Ilias y Moodle. Los productos de Blackboard, Catedr@, Desire2Learn Inc., eCollege y Saba Software son de pago, mientras que las plataformas Atutor, Chamilo LMS, Claroline, Docebo, Dokeos, Ilias, Moodle y Sakai CLE son de código abierto.

Moodle ha sido durante mucho tiempo un 'estándar de facto' como paquete de software libre para la creación de cursos y sitios Web basados en Internet, siendo utilizado por instituciones educativas estatales, como la Universidad del País Vasco / Euskal Herriko Unibertsitatea, para ofrecer formación e-learning y b-learning. Existe una gran cantidad de comparativas, informes y evaluaciones entre Moodle, Sakai, Blackboard y otros LMS realizado por varias organizaciones en el proceso de selección de una de estas plataformas, en las que Moodle resultó elegida en el 95% de los casos (Leyva, 2010). Según la información recopilada por el Instituto de Tecnologías Educativas (http://www.ite.educacion.es/) actualmente en España las plataformas de código abierto son las más utilizadas, ocupando el primer lugar Moodle seguida de Dokeos, Ilias y Claroline. Todas ellas han sido diseñadas en PHP/Apache/MySQL – el desarrollo es en el lenguaje de programación PHP, utilizan Apache como servidor Web y MySQL como sistema de gestión de bases de datos – y son multiplataforma. Todas soportan varios idiomas, entre ellos el castellano, y todas presentan las características y funcionalidades propias de los LCMS.

VII 2 Revisión detallada de la plataforma Moodle

Moodle basa su funcionamiento en *usuarios*, *cursos* y *módulos*. A su vez, los módulos se pueden clasificar en tres grupos (Büchner, 2011; Moodle, 2017; William, 2008): *recursos*, *actividades* y *bloques*.

En cuanto a los *usuarios* de la plataforma, cada uno tiene asociado un *perfil* formado por información personal en la que obligatoriamente deben aparecer sus datos de acceso a Moodle, un nombre, apellido(s), dirección de correo electrónico y ciudad. Moodle proporciona un servicio de autentificación propio, aunque también da soporte a mecanismos de autentificación externa. Los distintos tipos de usuarios se diferencian atendiendo a un *sistema de roles* que determinan sus posibilidades de interactuar con la plataforma. Cada rol asigna una serie de *privilegios* que permiten restringir ciertas

funcionalidades a un usuario. Cada rol existente en Moodle se identifica internamente mediante un número y, ordenados en función del nivel de privilegios de mayor a menor son: administrador principal (0), gestor (1), creador de curso (2), profesor (3), profesor sin derecho de edición (4), estudiante (5) e invitado (6). Estos roles se aplican organizados en contextos jerárquicos, es decir, un usuario puede ejercer un rol de alumno en un curso, pero ejercer el rol de administrador sobre una actividad dentro del curso. Los roles son heredables desde los niveles altos de la jerarquía. Así, si un usuario tiene el rol de profesor en un curso, también lo tendrá para todos los elementos dentro del mismo.

Concretamente, el rol de *administrador principal* (*Administrator*) permite gestionar/configurar toda la plataforma; el rol de *gestor* (*Manager*) permite acceder a cursos, modificarlos, y realizar algunas tareas a nivel administrativo relacionadas con cursos, usuarios, configuraciones de calificaciones, etc.; el rol de *creador de curso* (*Creator*) permite la creación de nuevos cursos; el rol de *profesor* (*Teacher*) tiene acceso completo para editar sus cursos y matricular a otros usuarios en ellos, puede añadir tanto recursos como actividades, cambiar configuraciones, introducir puntuaciones y configurar el expediente de puntuaciones; el rol de *profesor sin derecho de edición* (*Non editing teacher*) dentro de un curso puede ver y calificar el trabajo de los estudiantes, pero no puede alterar o eliminar ninguna de las actividades o recursos; el rol de *estudiante* (*Student*) permite participar en las actividades y descargar materiales del curso; y el rol de *invitado* (*Guest*) permite un acceso básico al curso, sin ingún tipo de privilegio adicional.

Los *cursos* constituyen su unidad básica de organización. Todos los cursos que se crean están clasificados en categorías. Dentro de cada curso se encuentran los recursos y/o las actividades. Los recursos están orientados a procesos de aprendizaje pasivos para cuestiones de auto-estudio, y priorizan la interacción persona-contenido (trabajos individuales con el contenido). Un *recurso* es un documento, imagen, página web, un enlace a un archivo, un enlace a una Web, una etiqueta, etc. que sirve como soporte a las actividades del curso. Una *actividad* se orienta a procesos de aprendizaje activos, y prioriza la interacción entre personas. Moodle ofrece gran cantidad de actividades como chats, foros, glosarios, blogs, wikis, sistemas de control de actividades de aprendizaje, talleres y cuestionarios. Un *cuestionario* es básicamente un examen de tipo test y no es más que un conjunto de preguntas, con la salvedad de que se incluye su puntuación. Las preguntas así definidas se almacenan en un *banco de preguntas* a nivel de curso y cada una de ellas es de alguno de los siguientes tipos: numérica, calculada, opción múltiple, opción múltiple calculada, respuesta corta, ensayo, emparejamiento, emparejamiento de respuesta corta aleatoria, verdadero/falso, descripción y respuestas incrustadas en formato Cloze. A nivel de cuestionario, permite configurar la puntuación final así como el modo de navegar por las preguntas y es capaz de dar realimentación (inmediata o diferida, por ítem o una vez concluido todo el cuestionario). Gracias a las puntuaciones definidas Moodle es capaz de calificar automáticamente los cuestionarios cumplimentados y permite visualizar tanto las respuestas dadas por cada usuario que intentó el cuestionario, como las *puntuaciones* obtenidas en cada apartado y los tiempos consumidos. Todos estos datos se pueden descargar directamente en *informes* de resultados con varios formatos como hoja de cálculo Excel u hoja de cálculo OpenOffice.

Antes de comenzar la impartición de un curso es necesario: (1) crear y configurar el curso, (2) crear y configurar las actividades del curso que el alumno

deberá realizar y (3) matricular a los participantes en el curso con el rol correspondiente. Una vez realizadas estas tareas hay que dar visibilidad al curso para que esté disponible para los alumnos. Las distintas opciones de configuración se encuentran en la última de las estructuras principales: los bloques. Los *bloques* de Moodle proporcionan información o funcionalidad adicional. Entre ellos, encontramos bloques de calendarios, de administración de un curso, de usuarios en línea, de novedades, de canales RSS remotos, etc. Los bloques aparecen en los laterales de las páginas de Moodle y puede configurarse, o no, su aparición incluso su apariencia. El más importante es el *bloque de Administración de Moodle*, que permite configurar el LMS y gestionar todos los servicios y herramientas disponibles como cursos, actividades, puntuaciones, informes, recursos, servicio de autentificación, usuarios, formato de la página principal, apariencia e idioma del sitio Web, servicios Web y mensajería. En lo referente a la *configuración del LMS*, el administrador puede tomar una serie de decisiones y configurar la plataforma en función del uso que se le vaya a dar. Esto significa que activará o no ciertas opciones disponibles en la plataforma y establecerá valores por defecto a nivel global, que después podrán ser cambiadas por otros usuarios en sus cursos siempre que su nivel de privilegios se lo permita.

Si se precisa que aplicaciones externas interoperen con Moodle, el administrador deberá activar los *servicios web de Moodle* que ofrecen las funcionalidades que lo posibilitan. Por ejemplo, existen servicios para crear cursos, crear, borrar y modificar usuarios, crear grupos de usuarios, subir archivos, recuperar los alumnos matriculados y matricular nuevos manualmente.

El administrador también puede configurar *propiedades por defecto para usuarios y roles*. Por ejemplo, decidir cuál será el método de autentificación, modificar los privilegios de algún rol, y personalizar los datos que se piden en el perfil de las cuentas de usuario. Si decide permitir el envío de mensajes entre distintos usuarios podrá *activar el servicio de mensajería*.

Además, puede *configurar el Look & Feel* de Moodle, decidiendo el formato de la página principal del sitio Web, la apariencia, el idioma, etc. Por ejemplo, el único idioma disponible en principio es el inglés, pero el administrador puede añadir más y establecer cualquiera de ellos como idioma por defecto.

Por otra parte, también es posible facilitar las labores del profesor configurando adecuadamente otros parámetros más específicos atendiendo a las necesidades más habituales de los usuarios potenciales de la plataforma. En este sentido, el administrador puede establecer una *configuración por defecto para los nuevos cursos*, especificando por ejemplo si el curso está visible para los alumnos o no y la estructura más probable para el curso (clasificando los contenidos por temas o por semanas). Si se debe controlar automáticamente si una actividad está disponible o no, puede *habilitar el rastreo del grado de finalización en el nivel curso y el acceso condicional*, que permiten fijar las condiciones – basadas en la fecha, la calificación o el grado de finalización – que lo controlan. Asímismo, el administrador puede plantear una *configuración por defecto para los nuevos cuestionarios* ajustando los parámetros que definen los valores por defecto usados en el formulario de especificaciones. También puede decidir qué *métodos de matriculación* estarán habilitados: manual (el profesor es el que da de alta a los alumnos en el curso) y/o auto-matriculación (el alumno se da de alta él mismo).

VIII Software para análisis de datos y calibración

Existen en el mercado múltiples herramientas software que automatizan la fase de análisis de datos y calibración de ítems en el parámetro dificultad, con distintos grados de cobertura para cada una de las tareas que la conforman. A la hora de abordar esta fase, el parámetro dificultad de cada ítem se estima llegando a un consenso entre los distintos expertos, para lo cual, como se vio en el capítulo VI , existen dos alternativas: (1) *volverlos a reunir* para que consensúen el nivel de dificultad de los ítems conflictivos o (2) *aplicar algún procedimiento estadístico* que lleve a cabo el consenso de forma automática. En el contexto de esta tesis doctoral la fase de análisis y calibración de ítems se lleva a cabo mediante la utilización de un procedimiento estadístico, por lo que este capítulo revisa exclusivamente este tipo de software. Esta compilación no es exhaustiva, pero sí incluye las herramientas más importantes o históricamente relevantes.

Existen distintas aplicaciones que soportan tareas concretas de la fase de análisis y calibración mediante el juicio de expertos utilizando procedimientos estadísticos. A nivel de metodología, estas tareas pueden estar basadas en la TRI o en la TCT.

La TRI proporciona innumerables métodos y modelos para las tareas del análisis psicométrico y de la calibración de un banco de ítems. En este sentido, para llevar a cabo la estimación estadística en el parámetro dificultad según los preceptos de la TRI, algo que queda más allá del alcance de esta tesis, habrán de seguirse los modelos de un parámetro, como el modelo de Rasch. Existe una variedad de *software que implementa los diferentes procedimientos existentes para llevar a cabo la estimación de parámetros* (Baker, 1992; López Pina, 1995). Se trata de programas estadísticos que calculan en pocos segundos las estimaciones de los parámetros mediante aproximaciones sucesivas, iterando hasta que los valores obtenidos convergen. El primer procedimiento implementado, el de estimación máximo verosímil conjunta (JML), se debe a Lord (1968), quien describió el programa pionero que evolucionó hasta convertirse en los disponibles actualmente. La estimación de parámetros no es la única fase de la calibración para la que existen programas específicos que automatizan sus tareas, pues también puede encontrarse *software de apoyo para las fases de análisis basados en la TRI*, como son los análisis de unidimensionalidad, el estudio de la bondad de ajuste o la equiparación de puntuaciones. De cara a efectuar la fase de análisis de unidimensionalidad se pueden utilizar modelos de ecuaciones estructurales (SEM, Structural Equation Models) o modelos de regresión. Hoy en día existen multitud de páginas web que recopilan este tipo de software, por ejemplo http://www.rasch.org/software.htm, que es un directorio de programas software de análisis que implementan el modelo de Rasch.

Por su parte, la TCT proporciona algunos métodos para analizar ítems y test basados en estadísticas simples como proporciones, correlaciones y medias. Sin

embargo, que los cálculos estadísticos empleados sean sencillos no significa que el análisis de un ítem sea fácil y en muchas ocasiones será preciso el uso del ordenador, y necesarias aplicaciones que soporten estos métodos. En este sentido, puede encontrarse *software de apoyo para los análisis clásicos de ítems y test propuestos por la TCT***.**

Por último, existen sistemas complejos y ambiciosos **que cubren todo el proceso de desarrollo de un test**, incluido el proceso de análisis y calibración de sus ítems mediante la TRI y, en algunos casos, también la TCT. Este tipo de software permite construir bancos de ítems, calibrarlos desde diversos modelos, efectuar análisis de fiabilidad, y administrar test informatizados tanto clásicos como adaptativos, dando soporte completo a los distintos modelos de la TRI. Además, pueden disponer de **paquetes independientes** que permiten realizar análisis y estimaciones concretas en nuestros propios procesos de calibración.

En esta compilación, para cada herramienta software considerada se indican las siguientes dimensiones: (1) el programa y su referencia, (2) qué cubre el programa dentro del análisis y calibración, (3) cómo lo cubre: métodos (3) evolución (4) sistema operativo soportado y (5) algunos ejemplos del uso que se le ha dado.

VIII 1 Revisión del software

Durante esta compilación se han buscado aplicaciones informáticas que pudieran dar una cobertura similar a la que da CALLIE al proceso de calibración, aunando los dos métodos de calibración (psicométrica y mediante juicios de expertos).

En los siguientes párrafos se detallan las distintas dimensiones para los programas revisados. Se citan siguiendo un orden alfabético: AMOS, ANCILLES, BICAL, BILOG, CITAS, EQS, FastTest, ITEMAN, LERTAP, LOGIST, METRIX, R, RASCAL, SIETTE y XCALIBRE.

AMOS (Arbuckle y Wothke, 1999) es un programa con gran potencialidad gráfica con el que resulta muy sencillo construir y utilizar modelos estructurales, por ejemplo de cara a efectuar estudios de unidimensionalidad. Este software, que puede integrarse en el paquete estadístico SPSS, extiende el análisis multivariado estándar e incluye regresiones, análisis factorial, correlaciones y análisis de varianza. Como características técnicas destaca que los datos a analizar pueden importarse desde varios formatos de bases de datos, tales como SPSS o desde un archivo Microsoft Excel. Como alternativa a la especificación de un modelo utilizando la interfaz gráfica del propio AMOS es posible especificar y ajustar un modelo mediante programación en VB.NET o en C#. Entre los usos que se le ha dado a AMOS, pueden citarse las pruebas de validez efectuadas sobre el banco de ítems de nivel de inglés que utiliza el actual sistema eCAT (Olea, Abad, Ponsoda y Ximénez, 2004) y la validación de la batería de TAI de inteligencia INSBAT, incluida en el Vienna Test System (Sommer, Arendasy y Häusler, 2005).

ANCILLES (Urry, 1978) fue el primer software en implementar un procedimiento heurístico para estimar los parámetros según el modelo 3PL. No obstante, este programa no se ha utilizado más que con fines comparativos porque dicho procedimiento no dispone de una base teórica bien cimentada. En particular,

ANCILLES fija el valor del parámetro *c* para después convertir los estadísticos tradicionales de la TCT en sus homólogos *a* y *b* de la TRI.

BICAL (Wright, Mead y Bell, 1979; Wright y Panchapakesan, 1969), que implementa la estimación máximo verosímil conjunta, sólo sirve para el modelo de Rasch. Pese a tratarse de uno de los programas de estimación de parámetros más antiguos, es uno de los más utilizados con frecuencia para modelar ítems de un parámetro. Se distribuye bajo los nombres comerciales BIGSCALE (Wright, Linacre y Schulz, 1989), que implementa los métodos de máxima verosimilitud incondicional y conjunta para ítems calibrados según el modelo de Rasch, MICROSCALE (Wright y Linacre, 1985), la versión para PC de BICAL desarrollada por la empresa Mediax Interactive Technologies, y, principalmente, BIGSTEPS (Linacre y Wright, 1997; Wright y Linacre, 1992) la variante más reciente, cuya versión para Windows se denomina WINSTEPS, que permite además estimar los parámetros de los ítems según el modelo de crédito parcial. Precisamente este último programa fue el elegido para calibrar el banco de ítems del examen NBOME de osteopatía según el modelo de Rasch (Shen, 1993) y los ítems del test GEPA de competencia para estudiantes de octavo grado según el modelo de crédito parcial (NJDoE, 2006), así como para obtener la medida multidimensional y analizar la bondad de ajuste de la prueba SODA de análisis de severidad de dispepsia (Rabeneck, Cook, Wristers, Souchek, Menke y Wray, 2001)**.**

BILOG (Mislevy y Bock, 1990), actualmente comercializado como BILOG-MG3, desarrollado por el ETS (Educational Testing Service de Estados Unidos), se considera como uno de los programas más completos para estimar los parámetros según cualquiera de los modelos dicotómicos de la TRI. Es útil para test cortos y muestras pequeñas. Este software implementa los métodos bayesianos EAP y MAP, así como la reformulación del método de máxima verosimilitud marginal en el contexto de los algoritmos EM de Bock y Aitkin (1981), y si se le proporciona una distribución previa de habilidades también calcula la estimación bayesiana marginal (Mislevy, 1986), asumiendo una distribución a priori normal para el parámetro de dificultad, una logístico-normal para la discriminación y una beta para el pseudoacierto (Swaminathan, Hambleton, Sireci, Xing y Rizavi, 2003). La primera versión de este software (Mislevy y Bock, 1984), comercializada bajo el nombre *BIMAIN* (Mislevy, Bock y Muraki, 1988) y pensada para trabajar en *mainframes*, permitía obtener tanto la estimación máximo verosímil marginal, evitando los ciclos EM mediante una aproximación rápida y no iterativa, como la bayesiana MAP para los ítems nuevos en el contexto de la calibración on-line. Poco después se desarrolló una versión para computadoras personales denominada *PC-BILOG* que incluía multitud de características nuevas y opciones de análisis (Mislevy y Bock, 1986), y que fueron mejoradas y ampliadas en las sucesivas versiones *BILOG3* (Mislevy y Bock, 1990) y *BILOG-MG* (Zimowski, Muraki, Mislevy y Bock, 1996). Esta última incluye procedimientos multigrupo que permiten efectuar estudios de funcionamiento diferencial de los ítems, y ofrece un estadístico para evaluar la bondad de ajuste de los ítems. El programa BILOG se caracteriza por utilizar un completo lenguaje de comandos tan versátil que permite al psicómetra experto afrontar prácticamente cualquier situación que se le presente a la hora de realizar la estimación de parámetros, incluidas la posibilidad de efectuar análisis de FDI y la calibración on-line de un banco de ítems. Para este caso, la herramienta toma, para cada ítem nuevo, la media y la varianza de los parámetros de los ya calibrados del banco como parámetros de la distribución a priori (Wainer y Mislevy, 2000). PC-BILOG se ha utilizado para calibrar y ajustar al modelo el banco de ítems de un test de selección de programadores

(Zickar, Overton, Taylor y Harms, 1999); BILOG para estimar los parámetros de pseudoacierto de un TAI multidimensional de lectura y matemáticas (Li y Schafer, 2005) y para analizar según los modelos 1PL, 2PL y 3PL los ítems del test de admisión para el colegio de médicos estadounidense MCAT (Hendrickson y Kolen, 1999); BILOG 3 para estimar y ajustar al modelo 3PL los parámetros de los ítems de las pruebas de conocimiento en el contexto de una red de escuelas (Marchesi, 2001); y BILOG-MG para obtener el modelo logístico de dos parámetros de un banco de ítems de medición y geometría para el que se efectuó un análisis de funcionamiento diferencial entre sus administraciones en inglés y francés (Emenogu y Childs, 2005).

CITAS (Classical Item and Test Analysis Spreadsheet) es un fichero Microsoft Excel u OpenOffice preparado para el análisis clásico de ítems y test. Es un sistema autocontenido, que está disponible de forma gratuita, y que proporciona todas las salidas esenciales para un análisis clásico de evaluaciones. Ha sido diseñado para ser una herramienta de análisis clásico fácil de utilizar para pequeños conjuntos de datos (hasta 50 ítems dicotómicos y 50 individuos), que permite a los no psicómetras medir la calidad de sus evaluaciones de una manera directa. Como programas profesionales de análisis clásico son más adecuados LERTAP e ITEMAN. No se han encontrado datos sobre su uso.

EQS – Structural Equation Modeling Software (Bentler, 1985; Bentler y Wu, 1993) es un paquete informático que permite realizar análisis de unidimensionalidad sobre modelos de relaciones estructurales, poniendo a prueba hipótesis complejas sobre su validez y controlando la influencia del error de medida en la estimación de los coeficientes. Podría decirse que este software sólo se ha utilizado en el contexto de la TCT, por ejemplo, para efectuar un análisis factorial confirmatorio en la medición del capital social (Narayan y Cassidy, 2001), verificar el modelo estructural de hipótesis en logística (Lynch, 2004), o realizar estudios comparativos entre la teoría clásica y la TRI (Singh, 2004), así como entre la TCT y los modelos cognitivos (Case, Demetriou, Platsidou y Kazi, 2001). Es un programa para Windows.

FastTest (http://www.assess.com), es un sistema que cubre todo el proceso de desarrollo de un test. Comercializado por la empresa estadounidense ASC - Assessment Systems Corporation (1979), su software está específicamente diseñado para organizaciones que utilizan test a nivel profesional. Fastest es el software más completo que existe actualmente en su ámbito, cubre todo el proceso de desarrollo de un test, permite entregar los test en múltiples formatos, proporciona acceso instantáneo a los resultados e informes de realimentación personalizados y análisis psicométrico. Este software permite construir bancos de ítems, calibrarlos desde diversos modelos, efectuar análisis de fiabilidad, y administrar test informatizados tanto clásicos como adaptativos con estrategias bayesianas y/o máximo verosímiles, dando soporte completo a los modelos dicotómicos de la TRI. Se comercializa mediante dos versiones del producto para PC y una para la Web: FastTest PC Test Development System para una creación de test y cuestionarios en formato papel; FastTest PRO que es la versión profesional para PC y permite desarrollar e implementar una gran variedad de test electrónicos (convencionales, aleatorios, adaptativos, etc.); y la versión online FastTest Web (http://www.fasttestweb.com). Para dar cobertura al análisis psicométrico se descompone en tres módulos: XCALIBRE, RASCAL e ITEMAN. Los dos primeros analizan y calibran mediante cualquiera de los tres modelos logísticos de la TRI: Rasch, 2PL y 3PL, mientras que ITEMAN analiza según la TCT. Dispone también de dos

utilidades para el análisis clásico: CITAS y LERTAP. En cuanto a las herramientas disponibles, Fastest dispone de todo tipo de herramientas administrativas y también de herramientas de análisis (según la TCT, la TRI o personalizadas) y de calibración según la TRI. En cuanto a la cobertura de estándares e-learning, da soporte a la entrega de test mediante el estándar IMS QTI. A pesar de su envergadura, se trata de una solución excesivamente rígida, resultando difícil de usar y de hacer compatible con otros módulos ya implementados (Prieto y Delgado, 1999).

ITEMAN (Thompson y Guyer, 2010) es un programa software diseñado para proporcionar informes profesionales sobre análisis de ítems y test utilizando la TCT. Dirigido a profesores, administradores de escuelas y diseñadores instruccionales incluso a los no expertos en psicometría, se caracteriza por disponer de una interfaz amigable que facilita la ejecución del programa, y por permitir añadir ítems y comentarios en los informes al utilizar el formato RTF como formato de salida. Actualmente, se comercializa como un módulo del sistema FastTest que complementa los programas RASCAL y XCALIBRE permitiendo estudiar la bondad de ajuste, pues calcula automáticamente para cada ítem tantas correlaciones biseriales como opciones de respuesta tenga. Ofrece además una serie de estadísticos descriptivos e índices que facilitan el análisis psicométrico de los ítems del banco, y es capaz de gestionar respuestas a ítems de elección múltiple y de escala Likert. ITEMAN fue el software elegido para realizar el análisis clásico de los ítems de, entre otros, el test LPCAT (Murphy, 2002), una prueba de selección de empleados (Lofgren, 2005) y el examen de licenciatura en enfermería utilizado en Taiwan (Lin, Tseng y Wu, 1999).

LERTAP (Laboratory of Educational Research and Test Analysis Package) es un fichero Microsoft Excel preparado para el análisis clásico de ítems y test utilizando la TCT. Se ejecuta como aplicación Microsoft Excel y produce una variedad de informes en forma de tablas junto a los gráficos relacionados. Existen versiones para Windows y para Macintosh. Cada informe es también una hoja de cálculo Excel. Se incluye una funcionalidad que permite la importación de archivos en el formato de texto predictivo iTAP. No se han encontrado datos sobre su uso.

LOGIST (Wingersky, 1983; Wingersky, Barton y Lord, 1982) implementa las estimaciones máximo verosímil conjunta e incondicional para los modelos logísticos de uno, dos y, con especial énfasis, tres parámetros. Además proporciona estimaciones equiparadas de los parámetros, incluso con respuestas omitidas, en cuyo caso modifica la función de verosimilitud; permite realizar una equiparación según el método de la calibración concurrente; e incorpora muchas opciones y excepciones de cara a prevenir errores e inconsistencias en los resultados, por lo que su uso no resulta sencillo si no se conocen los detalles del procedimiento de estimación máximo verosímil conjunta. Aun a pesar de los inconvenientes que este hecho puede plantear y las deficiencias detectadas en el procedimiento de máxima verosimilitud conjunta, LOGIST se ha convertido en el estándar de facto con el que se comparan los demás procedimientos de estimación de parámetros (Baker, 1992). A diferencia de BILOG, que, como ya se ha dicho, es útil para test cortos y muestras pequeñas, LOGIST es preferible cuando se utilizan test de más de 60 ítems y muestras numerosas, superiores al millar de sujetos (Baker, 1992). Este software se utilizó para estimar los tres parámetros de los ítems de la prueba CAT-ASVAB en su fase experimental (Wolfe, McBride y Sympson, 1997) o para caracterizar según el modelo 2PL los ítems de un test de aptitud musical (Vispoel, 1999).

METRIX (Renom, 1992) es un programa desarrollado en la Universidad de Barcelona (España) que funciona exclusivamente bajo el entorno Windows. Se comercializa en formato CD con la denominación *Metrix Engine* y es un potente módulo de análisis psicométrico de test psicológicos, cuestionarios, exámenes y encuestas, destinado a usuarios que deben tomar decisiones en base a este tipo de herramientas. El programa está diseñado combinando la sencillez de manejo con la facilidad de obtener rápidamente información detallada y gráfica sobre el potencial de los instrumentos de evaluación analizados. Su versión profesional admite por cada análisis un número ilimitado de examinados y 250 ítems agrupados en un máximo 10 pruebas diferentes, lo que le permite abordar proyectos de construcción de test y verificación de cualidades de instrumentos psicométricos a gran escala. Concretamente, *METRIX Engine* proporciona estimaciones para los modelos de un parámetro mediante la implementación del procedimiento de JML. Permite analizar las propiedades psicométricas de los ítems, ofreciendo diversos estadísticos descriptivos, índices y coeficientes para el estudio de ítems y escalas desde los puntos de vista de la TCT y la TRI, así como un análisis completo de distractores. METRIX fue el software elegido para realizar la estimación de parámetros y estudiar la fiabilidad durante el proceso de adaptación al español de la escala *Strategy Inventory for Language Learning*, que es la tradicionalmente utilizada en contextos anglosajones para la evaluación de las estrategias de aprendizaje usadas por los estudiantes de idiomas (Roncel Vega, 2007).

R (http://www.r-project.org/) es un lenguaje y entorno de programación para análisis estadístico y gráfico que se distribuye bajo la licencia GNU/GPL y está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux. Es uno de los lenguajes más utilizados en investigación por la comunidad estadística. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes. Su mayor fortaleza son los paquetes que tienen finalidades específicas y que extienden su configuración básica. Por ejemplo, hay grupos de paquetes relacionados con estadística bayesiana, econometría, series temporales, etc. Otros permiten utilizar R desde otras aplicaciones, por ejemplo para Microsoft Excel existe RExcel. En la actualidad existe un repositorio oficial, CRAN (https://ftp.cixug.es/CRAN/), con más de 6900 paquetes organizados en vistas que los agrupan según su naturaleza y función. La vista de tareas denominada "Métodos y Modelos Psicométricos" (o vista psychometrics) incluye todo lo concerniente tanto a la TRI como a la TCT, estimación, modelos, SEM, etc. Contiene múltiples métodos de estimación y múltiples modelos a los que aplicarlos: dicotómicos, politómicos y continuos. Esta vista se clasifica en seis categorías: Teoría de respuesta al ítem, Análisis de correspondencia, Modelos de ecuaciones estructurales, Escalado multidimensional, Teoría clásica del test y Análisis de la estructura del conocimiento. **A nivel de TCT**, tiene paquetes para realizar una gran variedad de tareas y análisis asociados con la TCT: (1) puntuar respuestas de selección múltiple, llevar a cabo análisis de fiabilidad, llevar a cabo análisis de ítems y transformar puntuaciones a diferentes escalas; (2) funciones para la teoría de correlación, meta-análisis (generalización de validez), fiabilidad, análisis clásico de ítems y utilidad; (3) funciones para comparar estadísticamente dos o más coeficientes alfa basadas tanto en grupos de individuos dependientes como independientes; (4) funciones para calcular el alfa de Cronbach, los coeficientes Kappa y los coeficientes de correlación intra-clase mediante distintos métodos (ICC); (5) un paquete que calcula y dibuja la curva de Cronbach - Mesbach , que es un método, basado en el coeficiente alfa de Cronbach de fiabilidad, para comprobar la unidimensionalidad de la escala de medición y (6) rutinas útiles en la psicología experimental y de personalidad. **A nivel de TRI** los paquetes incluyen:

aproximaciones bayesianas para estimar los parámetros de ítem y persona, así como distintas variaciones de la estimación de máxima verosimilitud (condicional, conjunta, marginal, media de Warm, etc.) o aproximación normal. Proporciona una interfaz común para la estimación de los parámetros del ítem en los modelos TRI para respuestas binarias con tres programas diferentes (ICL, BILOG-MG e Itm) y una variedad de funciones útiles en los modelos de la TRI. En concreto, para modelos dicotómicos, politómicos y continuos basados en el modelo de Rasch comprende: (1) el modelo de Rasch extendido (CMLE), es decir, para datos dicotómicos (RM), el modelo de test logístico lineal (LLTM), el modelo de crédito parcial (PCM), etc. (2) modelos de rasgo (habilidad) latente bajo la TRI (MMLE) (3) mezcla de modelos de Rasch y JMLE, (4) parámetros del modelo de Rasch mediante el algoritmo pairwise (PMLE), (5) modelos suplementarios de la TRI (PMLE, etc.) y (6) módulos de análisis de test (JMLE, MMLE) sobre modelos dicotómicos, politómicos, multifaceta, etc. R es un programa multiplataforma muy popular en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras. Por ello, R ha sido el software elegido para realizar diversos proyectos en ese campo. Sirvan como ejemplo Bioconductor en el campo del análisis de información genética, que comenzó en 2001 con el objetivo de desarrollar e integrar software para el análisis estadístico de datos de laboratorio en biología molecular; y Rmetrics, que comenzó en 2008, orientado al análisis de los mercados financieros y la valoración de instrumentos de inversión.

RASCAL (ASC, 1996) fue un programa independiente hasta su versión 3.5, pero hoy en día forma parte del sistema FastTest. RASCAL es un módulo para la calibración de ítems y puntuación de test, específicamente diseñado para el modelo de Rasch. Utiliza un procedimiento modificado de JML con un estimador modal bayesiano, incluye la posibilidad de estudiar la bondad de ajuste y permite realizar una equiparación según el método de la calibración concurrente. RASCAL además permite calibrar el parámetro dificultad de los ítems mediante el método de máxima verosimilitud incondicional, y con el objetivo de analizar la bondad del ajuste al modelo, utiliza el índice *Q*. RASCAL y su homólogo para modelos 2PL y 3PL – ASCAL (Vale y Gialluca, 1985) – se utilizaron, por ejemplo, para ajustar a la muestra los modelos de Rasch y logístico de tres parámetros durante el proceso de transición de un test de análisis lógico realizado en lápiz y papel a su versión adaptativa informatizada y basada en la generación automática de ítems (Revuelta y Ponsoda, 1998), para calibrar según el modelo 3PL bancos de ítems como el de vocabulario de inglés usado por ADTEST (Hontangas, Olea, Ponsoda, Revuelta y Wise, 2004) o los de un innovador test de mecanografía (Ho, Wang y Shyu, 1997), así como para efectuar los análisis del banco de ítems de LPCAT (Murphy, 2002). Una de sus características más útiles es la capacidad de ajustar las estimaciones de habilidad y del ítem que obtiene a las escalas utilizadas por otros programas.

SIETTE (Rios, Perez-de-la-Cruz y Conejo, 1998) es el acrónimo de Sistema Inteligente de Evaluación mediante Test para TeleEducación. Es un software dirigido al ámbito universitario, y desarrollado por el departamento de Lenguajes y Ciencias de la Computación de la universidad de Málaga. SIETTE (http://www.siette.org) es un sistema web que da cobertura a todo el proceso de desarrollo de un test tanto convencional como adaptativo: desde la creación de ítems y el almacenamiento de bancos de ítems hasta la realización de TAIs y de distintos análisis estadísticos según el modelo 3PL de la TRI (Conejo et al., 2004; García-Viñas, Conejo, Gastón, López y Ropero, 2013). Para ello, utiliza como marco teórico una implementación discreta del

modelo 3PL basada en la ecuación de Birnbaum (1968), lo que facilita enormemente la computación. Concretamente, en lugar de usar funciones en el dominio real para las curvas características, SIETTE emplea distribuciones finitas en las que la habilidad del alumno se estima mediante un valor discreto. La hipótesis en SIETTE es que el profesor ha calibrado la dificultad de las preguntas de forma no sesgada, es decir, se supone que el profesor ha cometido errores en la asignación del parámetro inicial de dificultad de las preguntas, pero que la distribución de estos errores sigue una normal cuya media es precisamente el valor de dificultad real, o lo que es lo mismo, la dificultad real de un conjunto de cuestiones de un test que se comportan frente a un conjunto de estudiantes como ítems de una misma dificultad, corresponde a la media de las dificultades estimadas asignadas inicialmente por el profesor. A un banco de preguntas que cumple esta condición le denominan conjunto de ítems equilibrado. La conclusión de todo ello es que si se acepta la hipótesis del conjunto equilibrado, solo se requiere un valor estimativo de la dificultad de las preguntas por parte del profesor. En cuanto a la clasificación del alumno en base a su habilidad, ésta se lleva a cabo utilizando la regla de Bayes. Respecto a los criterios de selección de preguntas y finalización del test también son adaptaciones de los criterios clásicos usados en la TRI con curvas características reales. Una de sus particularidades más relevantes es la posibilidad de realizar un test de forma colaborativa, entre varios estudiantes. En cuanto a las herramientas de que dispone, además de las propias de administración de test, posee herramientas para el análisis psicométrico. No da soporte a estándares e-learning. En cuanto al trabajo realizado con esta herramienta, destaca el desarrollo de un banco de ítems calibrado con más de 2500 preguntas sobre Botánica Forestal y la realización de más de 4300 sesiones de evaluación en la Escuela Universitaria de Ingeniería Técnica Forestal de la Universidad Politécnica de Madrid en colaboración con la Universidad de Málaga (García-Viñas et al., 2013). En la actualidad el sistema SIETTE está en vías de integrarse como módulo de evaluación dentro de un sistema tutor inteligente denominado MEDEA (MÉtodos De Enseñanza y Aprendizaje).

XCALIBRE (ASC, 2015), forma parte del módulo de análisis del sistema FastTest, aunque se puede utilizar de manera autónoma para calibrar bancos de ítems según los modelos de dos y tres parámetros. Para ello implementa el método máximo verosímil marginal sobre un algoritmo EM, asumiendo que la función de distribuciones previa $g(\theta)$ es una normal. Esta aplicación, que permite realizar análisis con un número ilimitado de respuestas a conjuntos de un máximo de 750 ítems, trabaja únicamente con modelos dicotómicos, aunque distingue entre aciertos, fallos, omisiones y no-respuestas. La diferencia entre estas dos últimas categorías es que, mientras que a los ítems omitidos se les asigna como probabilidad de acierto el inverso del número de posibles respuestas, las no-respuestas se excluyen sistemáticamente del análisis. XCALIBRE también permite estudiar la bondad de ajuste entre modelo e ítems, además de equiparar los parámetros obtenidos para los diferentes subtest según el método de la calibración concurrente o el media-sigma. Asimismo, ofrece la posibilidad de obtener las puntuaciones de los sujetos de la muestra mediante los métodos de máxima verosimilitud limitado al intervalo (-4, 4), MAP y EAP. Entre los múltiples usos que se le han dado a XCALIBRE se encuentran la calibración, siguiendo el modelo logístico de tres parámetros, de los bancos de ítems de la batería JPT (Brown y Brown, 2000), utilizada en la universidad de Hawaii para medir el nivel de japonés; de los ítems del test MBE (Thompson y Weiss, 2006), que se administra dos veces al año en los Estados Unidos para evaluar la habilidad de los candidatos a ejercer la abogacía cuando se trata de aplicar los principios legales en situaciones basadas en casos reales; de los ítems del

TAI piloto desarrollado en el King's College de Londres para evaluar a los estudiantes de primer curso de medicina sobre contenidos relativos a los sistemas cardiovascular y respiratorio (Heard, Byrne y Ward, 2002); y del banco de ítems del test de ingreso al sistema Hezinet (López-Cuadrado y Armendariz, 2006).

La Tabla 9 muestra la síntesis de los 15 programas estudiados y descritos en los párrafos anteriores.

Programa	Análisis TCT	Análisis TRI	Estimación TRI	Modelos TRI
AMOS		unidimensionalidad		
ANCILLES			Heurística	3PL
BICAL			IML, JML	Rasch
BILOG		bondad de ajuste equiparación (varios métodos)	bayesianos(EAP, MAP), MML, On-line	1PL, 2PL, 3PL
CITAS / FastTest	Sí			
EQS	Sí	unidimensionalidad		
ITEMAN / FastTest	Sí	bondad de ajuste		1PL
LERTAP / FastTest	Sí			
LOGIST		equiparación (concurrente)	IMV, JML	1PL, 2PL, 3PL
METRIX		bondad de ajuste	JML	1PL
R	Sí	unidimensionalidad bondad de ajuste equiparación	bayesianos(EAP, MAP), MML, On-line JML, etc.	Rasch 1PL, 2PL, 3PL
RASCAL		bondad de ajuste equiparación (concurrente)	JML, bayesiana	Rasch
SIETTE		unidimensionalidad bondad de ajuste equiparación	bayesiana	3PL
XCALIBRE		Bondad de ajuste Equiparación concurrente y media-sigma	MML	2PL, 3PL

Tabla 9 – Síntesis de programas de apoyo para el análisis y calibración de ítems.

Algunos programas que soportan la estimación del parámetro dificultad mediante el modelo logístico de un parámetro son RASCAL (siguiendo el modelo de Rasch), ITEMAN y METRIX. Los programas AMOS y EQS utilizan los modelos SEM para el análisis de unidimensionalidad. Los SEM son una familia de modelos estadísticos multivariantes – menos restrictivos que los modelos de regresión – que permiten estimar el efecto y las relaciones entre múltiples variables. Una de las ventajas más destacables de EQS es que su utilización no requiere conocimientos de álgebra matricial a diferencia de otros programas similares. XCALIBRE junto con ITEMAN constituyen un potente paquete de análisis de ítems y test para Windows. Solamente las herramientas R, CITAS y LERTAP permiten la implementación y posterior uso de procedimientos ad-hoc para el análisis de datos o la calibración.

Por otro lado, existe una serie de aplicaciones que son sistemas autocontenidos, es decir, el usuario registra los datos con los resultados del test (las respuestas del individuo junto con las respuestas correctas) y las especificaciones de los análisis en una o varias hojas de cálculo. Cuando selecciona las distintas opciones disponibles y se ejecutan, las estadísticas se actualizan en tiempo real y se crean los informes correspondientes en forma de tablas y gráficos. El propósito de estos informes es ayudar a los que utilizan test a evaluar la calidad de sus ítems y de los propios test, mediante el

examen de sus características psicométricas. Dos ejemplos, autocontenidos en Microsoft Excel, son CITAS y LERTAP.

A lo largo de esta revisión de software, las únicas soluciones encontradas que ofrecen una cobertura similar a la de CALLIE se corresponden con los dos sistemas que se han detallado aquí: SIETTE y los módulos ITEMAN, RASCAL y XCALIBRE de FastTest. Concretamente, SIETTE actualmente solo cubre modelos logísticos basados en la TRI con lo que no da cobertura a CALLIE-EXPERT. Por su parte, FastTest sí que cubre CALLIE-TRI (Armendariz, 2014) – que de hecho emplea resultados obtenidos con XCALIBRE –, pero no así CALLIE-EXPERT, al no contemplar la definición de procedimientos estadísticos ad-hoc. En consecuencia, ni el sistema SIETTE ni los módulos de calibración de FastTest permiten llevar a cabo el proceso de calibración propuesto en esta tesis. Por tanto, no se ha encontrado ninguna herramienta que, por sí sola, automatice todas las tareas de calibración como se propone en CALLIE.

Especialmente, no se ha encontrado ninguna aplicación informática que calibre de modo similar a CALLIE-EXPERT, esto es, que analice los datos recogidos a los expertos y estime la dificultad mediante métodos estadísticos propios basados en la TCT de forma totalmente automática. En función de las herramientas estudiadas, una posible solución de cara a esta automatización, consistiría en customizar los programas necesarios, utilizando sus entradas y sus resultados (en un mismo formato compatible entre ellos, por ejemplo ficheros Excel, y con los datos adecuados para cada momento) con objeto de cubrir la fase de análisis de datos y calibración del proceso global. Concretamente, como apoyo software para realizar los cálculos y análisis basados en la TCT (tanto análisis como estimación de la dificultad) propuestos para calibrar vía expertos en esta tesis se puede utilizar cualquier *paquete estadístico* del mercado, por ejemplo el programa *R* (http://www.r-project.org/), que permite tanto la importación de datos en Excel como la exportación de resultados en este mismo formato. Si se desea añadir más potencia gráfica a las estadísticas calculadas se puede emplear, como software de apoyo al proceso de análisis, de forma independiente o siguiendo el estilo de los sistemas autocontenidos, cualquier programa existente en el mercado que trabaje con hojas de cálculo, como pueden ser *Microsoft Excel*, o su equivalente en software libre *OpenOffice* (http://www.openoffice.org/).

Parte Cuarta: La herramienta de ayuda CALLIE-EXPERT

La Parte Cuarta condensa el trabajo realizado para diseñar e implementar **la herramienta de ayuda CALLIE-EXPERT**, una aplicación web que permite la calibración del parámetro dificultad de un banco de ítems utilizando el juicio de expertos.

El capítulo **IX - El sistema CALLIE-EXPERT** detalla la arquitectura y los metamodelos que utiliza la herramienta para automatizar el proceso de calibración vía expertos. Este capítulo incluye la descripción de la integración de la plataforma Moodle en el sistema, mediante la que CALLIE-EXPERT puede administrar los ítems a los expertos para su calibración.

El capítulo **X - Interfaz de CALLIE-EXPERT** se centra en las características generales de la interfaz de CALLIE y en la descripción detallada de las dos interfaces correspondientes a los módulos de CALLIE-EXPERT: CALLIE-ESKARI y CALLIE-PRO.

El capítulo **XI - Evaluación de CALLIE-EXPERT** detalla los resultados de los experimentos con usuarios, llevados a cabo para probar la idoneidad de cada uno de sus componentes y para realizar con CALLIE-EXPERT las calibraciones vía expertos de dos bancos de ítems con características de diseño muy dispares.

El capítulo **XII - Conclusiones** enumera las principales aportaciones que se han producido como consecuencia del trabajo documentado en los capítulos anteriores, describe posibles líneas futuras de trabajo que pueden dar continuidad a la presente tesis y hace un repaso por las publicaciones de la autora.

IX El sistema CALLIE-EXPERT

La herramienta CALLIE-EXPERT complementa a CALLIE-TRI (Armendariz, 2014) que automatiza las tareas del proceso de calibración psicométrico. CALLIE-EXPERT es una aplicación web estructurada siguiendo una arquitectura en tres capas que utiliza métodos de desarrollo dirigidos por modelos (MDDM). La lógica de negocio del sistema de calibración se centra en la especificación del proceso de calibración a realizar, en la descomposición en tareas del proceso, en la generación automática de otros artefactos software que permiten llevar a cabo dicho proceso y en la ejecución del propio proceso de calibración utilizando los elementos previamente generados. Con este objeto, el sistema se divide en tres componentes principales: la aplicación web principal, la plataforma educativa Moodle CALLIE-MOODLE y el motor de workflows de Microsoft WWF.

Como medio de comunicación y de persistencia de datos entre los distintos componentes del sistema se definen cinco metamodelos: tres que contienen la especificación completa del proceso de calibración y otros dos para el intercambio de datos y resultados.

En este capítulo se detallan los distintos metamodelos con los que trabaja CALLIE-EXPERT y la arquitectura del sistema. Concretamente, las dos primeras secciones se centran en la descripción de cada uno de los metamodelos: los tres que conforman la especificación de calibraciones de expertos (sección IX 1), y los dos metamodelos para la enumeración de las aportaciones realizadas por los expertos y el de los resultados obtenidos (sección IX 2). La última sección se dedica a la arquitectura de CALLIE-EXPERT y sus componentes (sección IX 3).

IX 1 Metamodelos para la especificación del proceso

CALLIE-EXPERT crea y utiliza un modelo en el que se presentan las distintas decisiones de acuerdo a un metamodelo de calibraciones, que se irá completando a medida que se vayan incluyendo las diferentes especificaciones que contempla.

Este metamodelo de calibraciones es el que muestra la Figura 9. Las peticiones de calibración (metaclase PeticiónCalibEXP) se recogen de diferentes responsables (metaclase Responsable). Cada petición debe incluir tres tipos de especificaciones, que se corresponden con otros tantos metamodelos: la especificación de la calibración (metaclase Calibración), la especificación del diseño de los cuestionarios (metaclase Diseño Cuestionarios) y la especificación del análisis y los cálculos (metaclase Análisis y Cálculos).

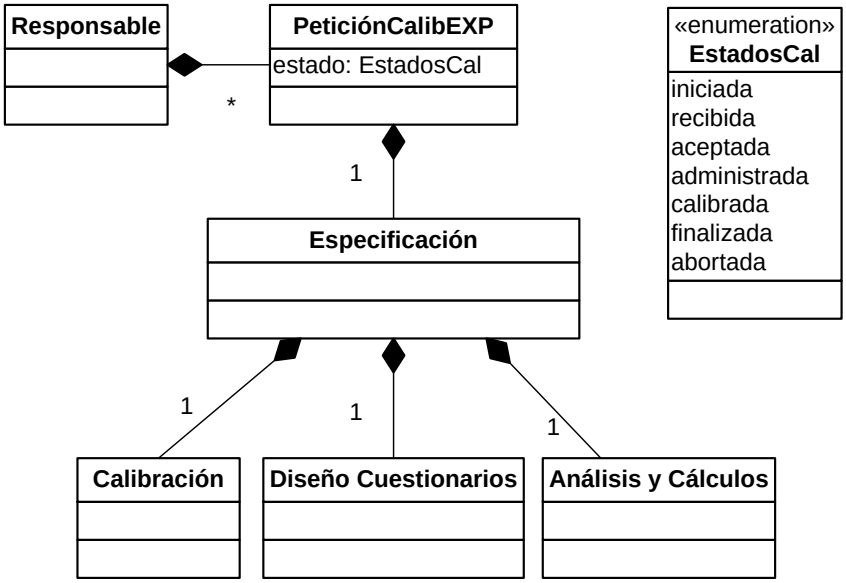


Figura 9 – Metamodelo de Calibraciones de CALLIE-EXPERT.

Cada calibración, determinada mediante una petición, tendrá un estado (metaclase Estado) con diferentes valores (metaclase enumerada EstadosCal).

Estos estados son inherentes al proceso de calibración asociado a cada petición que, desde que comienza hasta que termina, *pasa por una serie de estados* que se corresponden con las fases de especificación, administración, calibración y finalización del proceso dinámico asociado. Estas fases se realizan de manera sucesiva y es el responsable quien controla el progreso de una fase a otra siguiendo el diagrama de estados de la Figura 10.

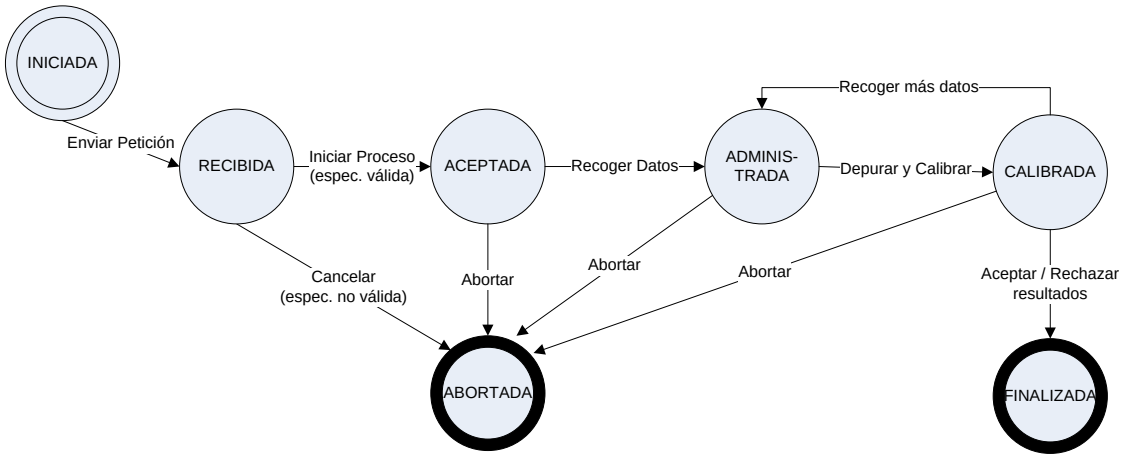


Figura 10 - Diagrama de estados del proceso de calibración en CALLIE-EXPERT.

Así, la calibración parte de un estado inicial, *iniciada*, que se produce cuando el responsable comienza a diseñar una nueva calibración y permanece en este estado hasta que completa su diseño. En el momento que el responsable envía la petición para su validación, el estado pasa a *recibida*. Una vez validado el modelo de calibración, en cuanto el responsable confirme el inicio del proceso, comienza la fase de administración, con la preparación de los artefactos necesarios para el proceso. En este

momento, el estado de la calibración cambia a *aceptada* y comienza la conducción del experimento. Cuando el responsable considere que se han rellenado suficientes cuestionarios, entonces el estado pasa a *administrada*. Permanece en él mientras se depura la muestra recogida y se realizan los cálculos de calibración. Una vez acabados estos cálculos el estado pasará a *calibrada*. Después de que la muestra haya sido depurada por primera vez el responsable de la calibración dispondrá de varias opciones a la vista de los resultados obtenidos por CALLIE-EXPERT. Dependiendo de su elección el estado de la calibración podrá volver a *administrada* (si el sistema descartó algunas administraciones y el responsable quiere incluir administraciones nuevas). Una vez que el responsable de la calibración acepta o rechaza los resultados obtenidos el estado de la calibración pasa a *finalizada*. Sin embargo, si el responsable decide cancelar el experimento en cualquier momento antes de su finalización, entonces el estado pasa a *abortada*.

El diagrama de estados de la Figura 10 también refleja las tareas que provocan las transiciones entre los distintos estados del proceso y que son: *Abortar* que pasa la petición a abortada, es decir, la cancela sin concluir; *Enviar Petición* que pasa la petición de iniciada a recibida; *Iniciar Proceso* que pasa la petición de recibida a aceptada; *Recoger Datos* que transcribe los datos recogidos de los expertos pasando la petición de aceptada a administrada, *Depurar y Calibrar* que pasa la petición de administrada a calibrada, así como *Aceptar/Rechazar Resultados* que pasa la petición de calibrada a finalizada. Mientras la petición está en fase de calibración, *Recoger (más) datos* indica la recogida de nuevas administraciones pasando la petición de calibrada a administrada. Dentro de un mismo proceso de calibración Enviar Petición, Iniciar Proceso, Aceptar/Rechazar Resultados y Abortar se producirán como máximo una sola vez, mientras que Recoger (más) Datos, y Depurar y Calibrar se podrán realizar de manera repetitiva hasta que el proceso termine.

Por último, CALLIE-EXPERT establece un plazo máximo de 6 meses para llevar a cabo todo el proceso de calibración, e implementa un evento interno denominado *Fin Experimento* que avisa al responsable si el proceso sobrepasa esa duración.

El objetivo de CALLIE-EXPERT siempre es calibrar ítems administrando cuestionarios a distintos expertos en la materia a la que pertenecen esos ítems. Por ello se especifica el tipo de calibración a realizar como "de expertos", se considera que el número de parámetros a obtener es 1 (la dificultad) y se indica un número mínimo de aportaciones a recoger por ítem (constante y preestablecido a 7). Para contener la especificación completa del proceso de calibración que se quiere llevar a cabo se definen tres metamodelos: el metamodelo de calibración (MMCA), que define el experimento de calibración, el metamodelo de los cuestionarios (MMCU), que define éstos y ciertos parámetros de administración de los mismos y un último metamodelo (MMANCA), que define los análisis y cálculos a realizar sobre las aportaciones que realicen los expertos.

IX 1.1 Metamodelo de calibración

El **metamodelo de calibración MMCA** especifica una serie de datos generales sobre la calibración concreta que se va a llevar a cabo (Figura 11) que amplía el metamodelo de la Figura 9. Los datos aportados tienen que ver con el objetivo del

experimento, el diseño inicial de los cuestionarios, el subsistema de administración a utilizar, los datos genéricos sobre el conjunto de expertos de que se dispondrá y la actuación del sistema de administración.

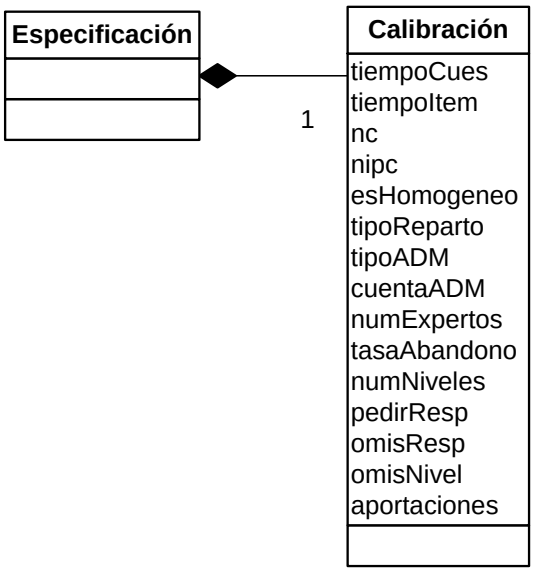


Figura 11 – Metamodelo de calibración MMCA.

Para llevar a cabo una preparación adecuada de los cuestionarios, existen dos datos importantes relacionados con el tiempo: la duración media de respuesta del ítem (*tiempoItem*) y la duración máxima por cuestionario (*tiempoCues*). Otros datos genéricos, necesarios también para el diseño inicial de los cuestionarios, se refieren a la cantidad de cuestionarios diferentes a diseñar (*nc*) y al número medio de ítems por cuestionario (*nipc*). Además, se debe especificar cómo va a ser el tamaño de estos cuestionarios: homogéneo o no (*esHomogéneo*), así como el modo inicial – alterno o continuo – con el que se asignarán los ítems a los distintos cuestionarios (*tipoReparto*). Por otro lado, CALLIE-EXPERT da la opción de utilizar un sistema de administración alternativo a Moodle, por lo que se debe indicar si la administración de los cuestionarios se hará a través de CALLIE-MOODLE o no (*tipoADM*). Para posibilitar la administración a los expertos con CALLIE-MOODLE, se deben indicar los datos (al menos nombre de usuario y contraseña) que permiten crear sus cuentas de acceso al curso en Moodle (*cuentaADM*).

Como datos genéricos sobre el conjunto de expertos se especifica la cantidad de expertos disponibles (*numExpertos*) junto a su previsión de abandonos (*tasaAbandono*) en la que se incluye también la tasa prevista de cuestionarios cumplimentados que no van a superar los filtros estipulados.

Además, para delimitar la actuación del subsistema de administración, se definen una serie de datos que podrán influir en el posterior análisis de la muestra recogida: el número de niveles de dificultad (*numNiveles*) que se manejan para calibrar los ítems; si en los datos a solicitar por ítem se pedirá a los expertos que lo resuelvan (*pedirResp*); si se permitirá dejar en blanco la respuesta al ítem y/o su nivel (*omisResp* y *omisNivel*); y si será posible que el experto haga comentarios globales sobre el cuestionario a cumplimentar (*aportaciones*).

IX 1.2 Metamodelo de cuestionarios

El **metamodelo de cuestionarios MMCU** especifica la configuración y diseño de los cuestionarios que serán administrados a los expertos en una petición de calibración y amplía el metamodelo de la Figura 9. Para ello, define los ítems a calibrar, cómo se van a estructurar estos ítems en los distintos cuestionarios y los expertos que van a participar en su calibración (Figura 12).

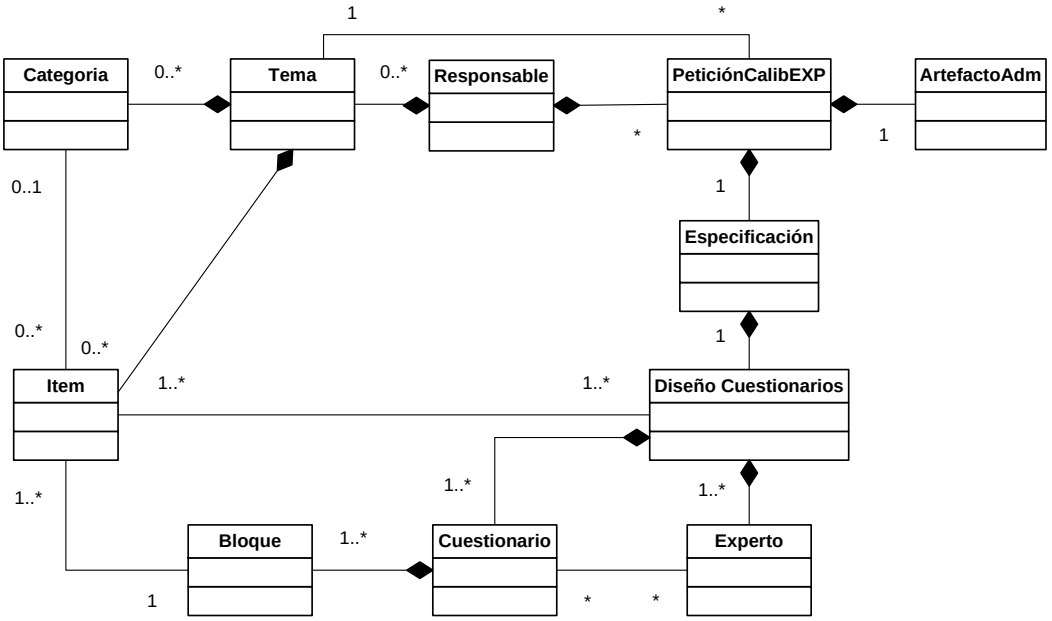


Figura 12 – Metamodelo de cuestionarios MMCU.

A cada petición de calibración se le asocia un artefacto para la administración de los ítems (metaclase *ArtefactoAdm*). En CALLIE-EXPERT este artefacto puede ser un curso Moodle de administración generado automáticamente a través de la creación previa de un paquete IMS. Cada objeto de esta clase guarda datos de configuración del propio artefacto tales como el formato, el idioma, las fechas en que estará disponible y si está visible o no para los expertos. También almacena datos de configuración para sus cuestionarios tales como el modo de navegación, la paginación, etc.

La especificación del diseño de los cuestionarios para una petición de calibración, y por tanto para el artefacto asociado, se almacena en una instancia de la metaclase *Diseño Cuestionarios* que contendrá la lista de ítems a calibrar, la lista de cuestionarios diseñados y la lista de expertos que participarán en la petición.

Los ítems a calibrar se especifican en formato IMS QTI (metaclase *Item*, que recoge todos los ítems objeto de la presente petición de calibración). Para facilitar las tareas del responsable, los ítems se deben clasificar en temas (metaclase Temas) y se ofrece la posibilidad de clasificar los ítems en subtemas denominados categorías (metaclase Categorías). Tanto los temas como las categorías los define el responsable. En cada petición solo se permite calibrar los ítems de un único tema, sin límite de categorías.

Cada cuestionario diseñado (metaclase *Cuestionario*) está formado por uno o varios bloques de ítems (metaclase *Bloque*). Se utiliza el concepto de *bloque* como el conjunto atómico de ítems distintos indivisible e inmodificable que forma parte de un cuestionario. En el caso de un diseño de cuestionarios con tamaños no homogéneos, cada bloque podrá formar parte de uno o más cuestionarios a administrar al experto; en caso contrario bloque y cuestionario serán equivalentes. Cada cuestionario diseñado está asignado a uno o varios expertos.

En una calibración basada en el juicio de expertos las contribuciones nunca son anónimas, puesto que el responsable debe conocer quién rellena cada cuestionario, por lo que todo participante debe estar identificado. Así, respecto a la información sobre el conjunto de expertos (metaclase *Experto*), de cada uno de ellos se guardan como mínimo sus datos de identificación y contacto electrónico (atributos *username*, *password*, *email* y *nombre*), a los que se pueden añadir datos extra que permitan al responsable determinar su perfil, establecer grupos de expertos o incluso agrupar resultados atendiendo a esas características (*edad*, *sexo*, *dirección*, *población*, *titulación* y *experiencia laboralen el área*).

IX 1.3 Metamodelo de análisis y cálculos

El **metamodelo de análisis y cálculos MMANCA** especifica todo lo que se quiere calcular y analizar en una petición de calibración y completa el metamodelo de la Figura 9. Los posibles análisis y cálculos a realizar que se deben especificar se encuentran en la biblioteca de análisis y cálculos de CALLIE-EXPERT.

La **biblioteca de análisis y cálculos** contiene análisis y criterios de filtrado que han sido descritos en el capítulo VI de esta memoria. Estos análisis y cálculos se clasifican en tres tipos: CriterioAdm, CriteriosCalibración y CalculosDificultad, que se pueden observar en el metamodelo de la Figura 13.

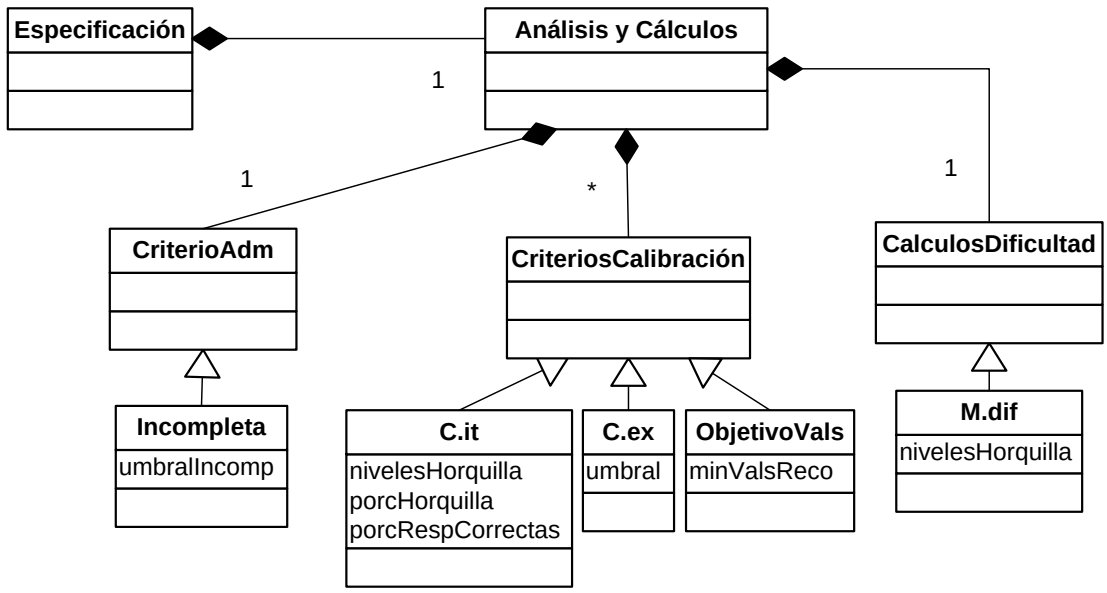


Figura 13 – Metamodelo de análisis y cálculos MMANCA.

El primer tipo (CriterioAdm) se corresponde con *análisis de administración* que utilizan criterios que no son específicos de una calibración mediante expertos. En este caso se ha añadido el análisis de administraciones incompletas, esto es, con menos de un determinado valor (*umbralIncomp*, que por defecto es 25%) de los ítems intentados (*Incompleta*).

Los otros dos tipos se corresponden con análisis y cálculos que se utilizan exclusivamente para la calibración mediante expertos. Dentro del tipo de análisis para calibración (CriteriosCalibración) se han añadido los criterios C.ex y C.it presentados en la sección II 3. Además, previendo los casos en los que para algún ítem no se llegue al mínimo de aportaciones de nivel válidas puesto como objetivo, se ha añadido un análisis adicional para garantizar que los estadísticos se aplican al objetivo recomendado: *ObjetivoVals*. En cuanto a los estadísticos para el cálculo de la dificultad (CalculosDificultad) se ha implementado el estadístico M.dif.

Mediante esta especificación, el responsable decide qué criterios utilizar en una calibración concreta y establece unos límites mínimos para que, cada vez que una administración o ítem no alcance esos mínimos, esa administración o ítem se invaliden junto con las aportaciones correspondientes.

Los parámetros de la Figura 13 son los que se deben concretar a la hora de especificar el análisis o estudio a realizar (por ej. C.it-1 con un 25% de respuestas correctas). A la hora de instanciar estos criterios existen tres valores: un objetivo de valoraciones por ítem (*MinValsReco* por defecto 7), un umbral para descartar las administraciones incorrectas (*umbral* por defecto 75%) y otro para descartar las administraciones incompletas (*umbralIncomp* por defecto 25%). Respecto al resto, serán los que deben especificarse en la petición de calibración. Así, si el responsable desea estudiar la fiabilidad de los ítems aplicando los criterios C.it debe especificar *porcRespCorrectas* para C.it-1, y *nivelesHorquilla* y *porcHorquilla* para C.it-2; si desea filtrar administraciones debe especificar si se tratarán las incorrectas, las incompletas o ambas.

IX 2 Metamodelos para el intercambio de datos y resultados

Además de los metamodelos que contienen la especificación, se han definido dos metamodelos más para el intercambio de datos y resultados: el metamodelo de aportaciones (MMAP), que describe los resultados que deben provenir de cada uno de los expertos y el metamodelo de resultados (MMRE), que define cómo serán los resultados que se van a obtener a partir de los análisis y cálculos realizados.

IX 2.1 Metamodelo de aportaciones

El metamodelo de aportaciones MMAP define el formato de los datos que se recogen en una petición de calibración. Los modelos correspondientes sirven para

alimentar al workflow de análisis y cálculos. El metamodelo MMAP incluye los datos suministrados por los expertos como queda representado en la Figura 14.

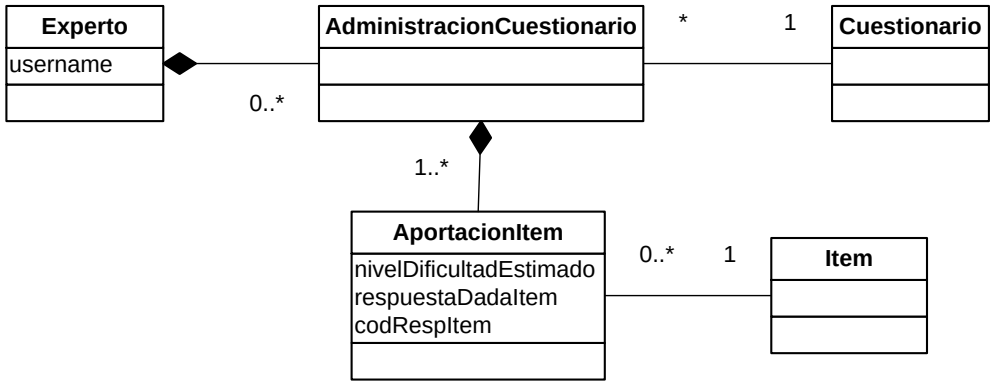


Figura 14 – Metamodelo de aportaciones MMAP.

Por cada experto participante en la petición de calibración (metaclase Experto) se almacenan los datos correspondientes a cada cuestionario que se le ha administrado. Así, por cada cuestionario administrado a un experto, existe una instancia de AdministracionCuestionario en la que se registran las respuestas de ese experto sobre ese cuestionario. Cada instancia de AportacionItem contiene los datos aportados por el experto para un ítem de esa administración, concretamente el nivel y la respuesta al ítem (atributos *nivelDificultadEstimado* y *respuestaDadaItem*). Además, para su posterior tratamiento, cada respuesta dada se guarda codificada como incorrecta, correcta o en blanco (atributo *CodRespItem*).

Cada cuestionario tendrá tantas administraciones (metaclase *AdministraciónCuestionario*) como expertos respondan a sus ítems. Por cada cuestionario diseñado (metaclase Cuestionario) existirán varias administraciones y por cada ítem a calibrar (metaclase Item) existirá un conjunto de aportaciones sobre ese ítem, desde ninguna a varias.

Debido a las posibilidades de modificación que ofrece CALLIE-EXPERT puede haberse eliminado un experto, un ítem o una de sus administraciones, por lo que puede no haber administraciones de un experto o aportaciones sobre un ítem concreto.

Además, MMAP solamente recoge los datos relevantes para las actividades de análisis y calibración tal y como las concibe CALLIE-EXPERT. Así, almacena exclusivamente el último intento del experto puesto que es el que acumula las aportaciones definitivas. Para el procesamiento de los datos recogidos es suficiente con que cada experto quede identificado mediante su nombre de usuario (atributo *username*) pudiéndose obviar en este metamodelo tanto sus datos personales como las aportaciones propias. Por tanto, no se transcriben las informaciones innecesarias para los cálculos, tales como si se pidió algún otro campo para cada ítem aparte de la respuesta y el nivel, ni las posibles aportaciones globales ni los intentos intermedios del experto.

El metamodelo MMAP permite recoger y analizar los datos aportados por los expertos, no solamente desde CALLIE-MOODLE, sino desde cualquier otro sistema de administración de cuestionarios, ya que puede representar tanto una base de datos como cualquier otro tipo de almacenamiento, por ejemplo un fichero Excel.

IX 2.2 Metamodelo de resultados

El metamodelo de resultados MMRE (Figura 15) almacena los datos detallados del análisis de la muestra recogida conforme al MMAP y del calibrado de los ítems y permite al responsable de la calibración visualizar y/o recuperar tanto los cálculos automáticos intermedios como el resultado de la calibración realizada por el sistema CALLIE-EXPERT. Con esta información el responsable podrá decidir con criterio si acepta o no dichos resultados. El MMRE ofrece un resultado por cada criterio del MMANCA aplicado.

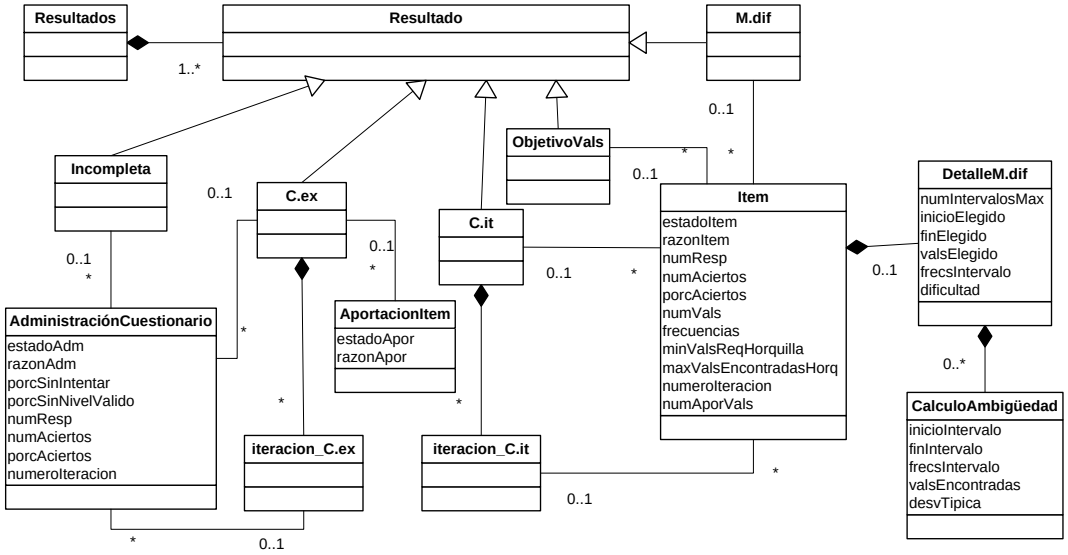


Figura 15 – Metamodelo de resultados MMRE.

La ejecución de los análisis Incompleta y C.ex del MMANCA tiene unos **resultados sobre la fiabilidad del experto** en cada una de sus administraciones que se reflejan en las metaclases *AdministracionCuestionario* y *AportacionItem* del metamodelo MMAP. Por un lado, el análisis de administraciones mediante el criterio **Incompleta** obtiene una lista de expertos válidos y rechazados, hallando el porcentaje de ítems del cuestionario sin intentar de cada administración (*porcSinIntentar*). Por otro lado, el análisis de administraciones mediante el criterio **C.ex** también obtiene una lista de expertos válidos y rechazados, y además se puede aplicar de forma iterativa, con lo que también se puede obtener una lista con los resultados de cada una de estas iteraciones (metaclase *Iteracion_C.ex*). Tras la aplicación de cualquiera de estos dos análisis, toda administración de un cuestionario a un experto (metaclase *AdministracionCuestionario*) obtiene como resultado un estado final (atributo *estadoAdm*: válida o rechazada) junto con el nombre del criterio concreto que ha causado su rechazo (atributo *razonAdm*). Además, la aplicación de estos criterios obtiene unos resultados detallados para cada una de las administraciones analizadas: el porcentaje de aportaciones sin nivel válido (*porcSinNivelValido*), el número de ítems que ha respondido el experto en ese cuestionario (*numResp*) junto con el total de aciertos (*numAciertos*) y el porcentaje que representa (*porcAciertos*) indicando las iteraciones realizadas hasta su descarte o aceptación final (*numeroIteracion*). Por

último, la aplicación del análisis C.ex genera también una lista de resultados individuales por aportación, que se reflejan en la metaclase *AportacionItem* del metamodelo MMAP.

La ejecución del análisis C.it de MMANCA tiene unos **resultados sobre la fiabilidad del ítem** que se reflejan en la metaclase *Item* del metamodelo MMAP. El análisis de ítems mediante el criterio **C.it** obtiene una lista de ítems válidos y descartados. Además, C.it se puede aplicar de forma iterativa con lo que también se genera una lista con los resultados de cada una de estas iteraciones (metaclase *Iteracion_C.it*). Tras la aplicación de este análisis, cada ítem a calibrar (metaclase *Item*) obtiene un estado (atributo *estadoItem*: válido o rechazado) junto con el nombre del criterio concreto que ha causado su rechazo (atributo *razonItem*). La ejecución de M.dif sobre los ítems no descartados por estos filtros (*estadoItem* válido) obtendrá su dificultad. Además, la aplicación de este criterio obtiene unos resultados detallados para cada uno de los ítems analizados: el número de respuestas dadas a ese ítem (*numResp*), el total de aciertos (*numAciertos*) y el porcentaje de respuestas correctas que representa (*porcAciertos*), la indicación de si ese ítem ha superado el análisis o no, el número de aportaciones que se mantienen (*numVals*), el número de estimaciones para cada nivel (*frecuencias*), el mínimo de valoraciones necesarias para cumplir el porcentaje en horquilla (*minValsReqHorq*) y el máximo encontrado (*maxValsEncontradasHorq*) indicando las iteraciones realizadas hasta su descarte o aceptación final (*numeroIteracion*).

La ejecución del análisis ObjetivoVals de MMANCA tiene unos **resultados sobre la fiabilidad de los resultados obtenidos**, que se reflejarán en la metaclase *Item* del metamodelo MMAP. El análisis de ítems mediante el criterio **ObjetivoVals** obtiene una lista de ítems válidos y sospechosos. Tras la aplicación de este análisis, cada ítem a calibrar (metaclase *Item*) obtiene un estado (atributo *estadoItem*): rechazado, válido o sospechoso junto con el nombre del criterio correspondiente en el caso de que no sea válido (atributo *razon*). Además, la aplicación de este criterio obtiene unos resultados detallados para cada uno de los ítems analizados: el número de valoraciones válidas recabadas para el ítem (*numAporValidas*).

El cálculo de las dificultades de los ítems mediante el procedimiento M.dif del MMANCA, obtiene un **resultado para el parámetro dificultad** que se refleja en la metaclase *Item* del metamodelo MMAP. Concretamente, la aplicación de **M.dif** genera como resultado una lista de ítems calibrados en dificultad, en la que para cada ítem se obtiene el número real que indica el nivel de dificultad al que pertenece (atributo *dificultad* de *DetallesM.dif*). Además, la aplicación de M.dif obtiene unos resultados detallados para cada uno de los ítems calibrados: los cálculos correspondientes a M.dif-1 (resto de atributos de la metaclase *DetallesM.dif*) y a M.dif-2 en caso de que sea aplicable (metaclase *CálculoAmbigüedad*). Concretamente, se almacenan todos los cálculos intermedios (frecuencias, valoraciones requeridas y encontradas, etc.) necesarios para estimar la dificultad de cada ítem del banco que no ha sido rechazado, incluidos los referidos a los ítems con intervalos ambiguos junto con sus desviaciones típicas y los intervalos finalmente elegidos.

IX 3 Arquitectura de CALLIE-EXPERT

El sistema CALLIE para la calibración de bancos de ítems basada en juicios de expertos se divide en tres componentes principales (Figura 16): CALLIE-MOODLE, que es el sistema para recoger las opiniones de los expertos basado la plataforma educativa Moodle, WWF que es el motor de flujos de trabajo de Microsoft y CALLIE-EXPERT que es la aplicación Web principal para especificar y realizar los análisis y la calibración. El componente CALLIE-EXPERT consta, a su vez, de cuatro módulos: ESKARI, ADMINQ Factory, WF Factory y PRO-C. Además, cada proceso de calibración se asocia a una *petición de calibración* y la generación automática consiste en crear tres elementos por cada petición: un *modelo de calibración*, un *curso y cuentas* de usuario para la administración de los cuestionarios en Moodle y un *workflow CA* con los distintos cálculos y análisis del proceso. Todo el proceso de calibración será diseñado y controlado por el responsable de la calibración a través de la interfaz de la aplicación, presente en los módulos ESKARI y PRO-C.

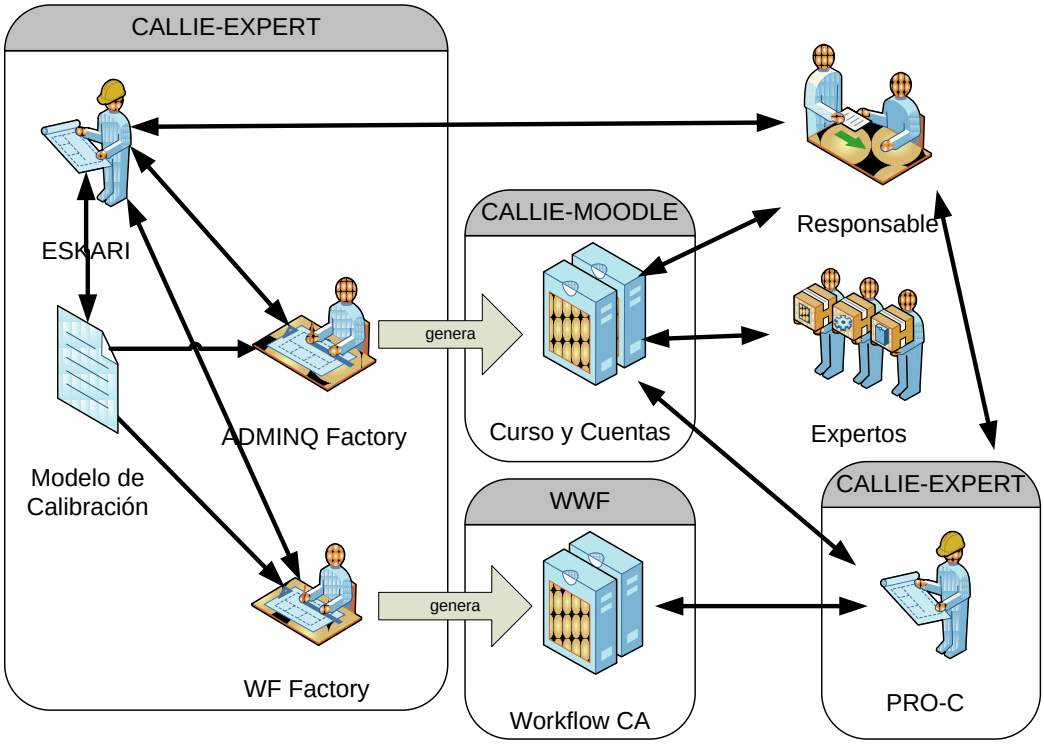


Figura 16 – Arquitectura del sistema de calibración.

El módulo ESKARI de CALLIE-EXPERT, se encarga de construir el modelo completo de calibración a realizar. A partir de ese modelo de calibración, el sistema define los cuestionarios a administrar y genera el artefacto (Curso y Cuentas) que se encargará de realizar las administraciones, mediante ADMINQ Factory y en colaboración con el sistema CALLIE-MOODLE. También, mediante WF Factory se genera un workflow con las tareas a realizar para llevar a cabo los análisis y cálculos necesarios para la calibración: el Workflow CA. Este workflow, se ejecutará a través del

motor de workflows WWF. Finalmente, el módulo PRO-C de CALLIE-EXPERT permite supervisar el proceso de administración de los cuestionarios, anular manualmente administraciones y dar por finalizada la recogida de aportaciones, dando paso a la realización de los cálculos que debe llevar a cabo el workflow antes mencionado. En los siguientes apartados se detallan en profundidad todos estos componentes.

IX 3.1 CALLIE-EXPERT: ESKARI

ESKARI interactúa con el responsable de la calibración, mediante su interfaz CALLIE-ESKARI, de la que se hablará en la sección X 1, y le permite diseñar y confirmar peticiones de calibración.

ESKARI es el encargado de crear los modelos conformes a MMCA, MMCU y MMANCA, de acuerdo con las indicaciones del responsable. La unión de estos tres modelos conforma el *modelo de calibración*, que contiene la especificación completa del proceso de calibración que se quiere llevar a cabo. ESKARI también se encarga de solicitar a ADMINQ Factory y a WF Factory la creación de los artefactos software necesarios objeto de su cometido. Para ello, ESKARI dispone de varias funcionalidades:

(1) *Diseñar Petición*. Solicita al responsable de la calibración todos los datos correspondientes a la tarea de diseño del experimento, necesarios para crear los modelos.

(2) *Enviar Petición*. Registra las especificaciones dadas, las valida y crea el modelo de calibración correspondiente.

(3) *Iniciar Proceso*. Invoca a ADMINQ Factory, que genera el curso y las cuentas en Moodle correspondientes a esa calibración si es necesario, e invoca a WF Factory que prepara el workflow CA.

Además, CALLIE-EXPERT dispone de una serie de *valores por defecto* que son los recomendados por la propia aplicación para cada uno de los metamodelos involucrados. Con estos valores, ESKARI puede calcular todos los datos necesarios y crear automáticamente un modelo de calibración acorde a la petición del responsable.

En primer lugar, se describen los **valores por defecto** que utiliza CALLIE-EXPERT para crear automáticamente el **modelo conforme a MMCA**. Respecto a las duraciones estimadas, el sistema parte del hecho de que cada ítem del banco es un ítem de selección múltiple – con 6 opciones como máximo – que se contesta en aproximadamente 1 minuto de media. Por otro lado, para estimar la duración total del cuestionario influyen diversos factores relacionados con el perfil del experto, principalmente con su motivación y cansancio. El sistema prevé que un experto sea capaz de permanecer como máximo una hora haciendo un cuestionario por lo que realiza de forma automática los cálculos para una duración total del cuestionario de 1 hora – duración corta para evitar abandonos por cansancio, desmotivación del experto – de donde 15 minutos son para leer las instrucciones de cumplimentación del cuestionario y para indicar los datos personales del experto (si se requieren) y los 45 minutos restantes se dedicarán a rellenar los ítems suponiendo que cada ítem tarda en cumplimentarse 1 minuto de media. Como subsistema de administración utilizará CALLIE-MOODLE y proporcionará un usario y contraseña a cada experto involucrado. En cuanto a las decisiones sobre el formato de los cuestionarios, los valores por defecto

de CALLIE-EXPERT son tamaño homogéneo en los cuestionarios, reparto inicial alterno, un único cuestionario con todos los ítems a calibrar y sin tasa de abandono por parte de los expertos, por tanto, 7 expertos para conseguir el mínimo de valoraciones por ítem. Estimar la dificultad del ítem en una escala de 10 niveles posibles, pedir y obligar a dar la respuesta correcta al ítem, obligar a estimar el nivel y permitir aportaciones globales.

Una vez establecido el banco de ítems a calibrar, CALLIE-EXPERT puede completar un modelo conforme al MMCU utilizando los valores recomendados por la propia aplicación. A continuación se describen los **valores por defecto** que utiliza CALLIE-EXPERT para crear automáticamente este **modelo conforme a MMCU**. En cuanto al diseño de cuestionarios, el sistema utiliza los valores por defecto del MMCA por lo que reparte uniformemente todos los ítems en distintos cuestionarios respetando las duraciones estimadas. Al asumir un diseño homogéneo, cuestionario y bloque coinciden. Para calcular los valores por defecto del número de cuestionarios a diseñar establece un *número inicial de ítems por cuestionario*, un número aproximado – que se calcula a partir de las duraciones estimadas – y que se transforma en el número entero igual o inferior más próximo que garantiza el diseño de unos cuestionarios homogéneos. Este número, junto con la cantidad de ítems a calibrar, sirve para conseguir una cantidad homogénea de ítems por cuestionario, es decir, los valores por defecto para el *número de cuestionarios* y para el *número de ítems por cuestionario* a partir de una aproximación por exceso. Por otro lado, para alcanzar el objetivo de valoraciones por ítem indicado en el MMCA el sistema asume en sus cálculos por defecto que se podrán captar todos los expertos necesarios y que no habrá tasa de abandono ni se descartará a ninguno de ellos. Por tanto, el *número de expertos disponibles* por defecto será siete veces el número de cuestionarios por defecto. Con estos valores calculados inicial y automáticamente por el sistema, CALLIE-EXPERT crea una asignación de partida de los ítems del banco en los distintos cuestionarios utilizando el reparto inicial alterno.

Para poder completar el modelo MMCU, el sistema crea automáticamente para cada participante experto datos identificativos y de contacto, concretamente una cuenta de acceso en CALLIE-MOODLE de la forma expertusXXX / Exp.123456, donde XXX es un número distinto para cada cuenta, a la que añade un email y un nombre ficticios que se podrán modificar posteriormente. Por último, asigna los cuestionarios a los expertos asumiendo que los cuestionarios diseñados se administrarán de manera uniforme entre los distintos expertos disponibles.

Los cálculos de valores recomendados también posibilitan a CALLIE-EXPERT llevar a cabo una serie de comprobaciones para detectar si el diseño de los cuestionarios que especifica el responsable cumple las restricciones impuestas por el sistema.

En último lugar, se describen los **valores por defecto** que utiliza CALLIE-EXPERT para crear automáticamente el **modelo conforme a MMANCA**. Estos valores se corresponden con los recomendados en la calibración del banco de ítems que se llevó a cabo para el sistema Hezinet (Arruabarrena, 2010) y son: no descartar las administraciones incompletas, realizar el estudio de fiabilidad del experto aplicando los criterios C.ex y estudiar la fiabilidad de los ítems aplicando los criterios C.it, eliminar ítems cuyo porcentaje de respuestas correctas sea inferior al 70%, establecer como horquilla el número impar más próximo al 35% del número de niveles y la malla más estricta del 85% de estimaciones dentro de esa horquilla.

IX 3.2 CALLIE-EXPERT: ADMINQ Factory

ADMINQ Factory crea un artefacto, conforme al estándar IMS, para la administración de cuestionarios de acuerdo con los modelos conformes a MMCU y a MMCA que haya creado ESKARI. A continuación, traslada ese artefacto a un LMS Moodle para que realice la administración y almacene las aportaciones realizadas por los expertos, que PRO-C recogerá en un modelo de acuerdo al MMAP. Este traslado a Moodle crea automáticamente un *curso en la plataforma Moodle* con una configuración inicial, junto con las *cuentas de acceso al curso* de los distintos participantes, para la administración electrónica mediante Web de los ítems especificados en una determinada calibración.

Para automatizar la creación de un curso conforme al estándar IMS como se define en los modelos conformes a MMCU y MMCA, ADMINQ Factory utiliza una *herramienta de generación de paquetes IMS*. Esta herramienta es una aplicación independiente que genera paquetes IMS a partir de las especificaciones de los ítems, de las decisiones de diseño de cuestionarios tomadas para la calibración y del metamodelo UML de la especificación IMS utilizando MDDM. Concretamente, se utiliza el paradigma M2C (Model to Code) para generar el fichero comprimido final, los distintos ficheros y carpetas del paquete y el paradigma M2T (Model to Text) para los distintos modelos XML que contiene el paquete. Se ha utilizado la herramienta proporcionada por IMS en su Web para validar el cartucho IMS CC generado. La aplicación se ha incluido como biblioteca ejecutable en CALLIE. Esta herramienta de generación de paquetes IMS puede crear tanto paquetes de contenido IMS CP como cartuchos IMS CC, siguiendo los modelos vistos en el capítulo IV de la presente memoria, y es capaz de generar cuestionarios en todas las versiones de IMS QTI indicadas en ese capítulo.

En cuanto a su comportamiento, ADMINQ Factory especifica el curso y sus cuestionarios utilizando la herramienta de generación de paquetes IMS según las especificaciones presentes en el modelo conforme a MMCU y a continuación, invoca a distintos servicios Web de CALLIE-MOODLE para el resto de tareas de configuración y creación. Además, ADMINQ Factory mantiene cierta información del sistema que posibilita crear nuevas cuentas de acceso (usuario y contraseña) en Moodle para los expertos involucrados, llevar el recuento de las ya utilizadas y recoger los datos desde Moodle de manera automática. También realiza una preasignación de expertos a los distintos cuestionarios conforme a los datos del MMCU.

En primer lugar, ADMINQ Factory crea y despliega, para cada petición de calibración, un curso de administración conforme al estándar IMS. Además, el curso creado se carga con los cuestionarios que se van a administrar a los expertos. Opcionalmente, cada experto puede realizar aportaciones a su colaboración de un modo global. ADMINQ Factory preconfigura el curso y los cuestionarios con opciones que facilitan al responsable las tareas de administración de cuestionarios a expertos. Concretamente configura el curso en el momento de su creación con un formato de un único tema, idioma castellano y disponible para sus expertos. Por otra parte, ADMINQ Factory da una misma configuración a todos los cuestionarios del curso. El cuestionario se configura de manera que el experto pueda completarlo en varios intentos, de modo que en cada nuevo intento se conserve lo ya contestado anteriormente, con navegación no lineal, sin límites de tiempo, sin realimentación sobre cuáles son las respuestas correctas y con todos los datos que se piden por cada ítem a calibrar en una misma

pantalla. El responsable podrá adaptar estos parámetros a su conveniencia si fuera necesario. La Figura 17 muestra un curso creado en Moodle con los distintos cuestionarios y un espacio para aportaciones propias.

Figura 17 – Página de Moodle correspondiente a un curso de administración.

Por cada ítem a calibrar, el experto debe aportar al menos dos datos: el nivel de dificultad estimado (dato obligatorio) y la respuesta correcta del ítem (dato opcional). Obviamente, cuando se solicita la respuesta correcta, el experto puede indicar en el mismo cuestionario cuál es la opción correcta de dicho ítem. El ítem a calibrar tendrá su propio diseño (dependiendo del tipo de ítem que sea), mientras que para el nivel de dificultad existen varias opciones: caja de texto o selección múltiple. La Figura 18 ilustra el formato de presentación correspondiente a un cuestionario llamado "Cuestionario_1" compuesto por 42 ítems a calibrar y muestra los datos que se piden para el ítem denominado EUSK1 (supuestos 12 niveles posibles de dificultad y presentación en forma de selección múltiple vertical para responder al ítem y horizontal para estimar el nivel).

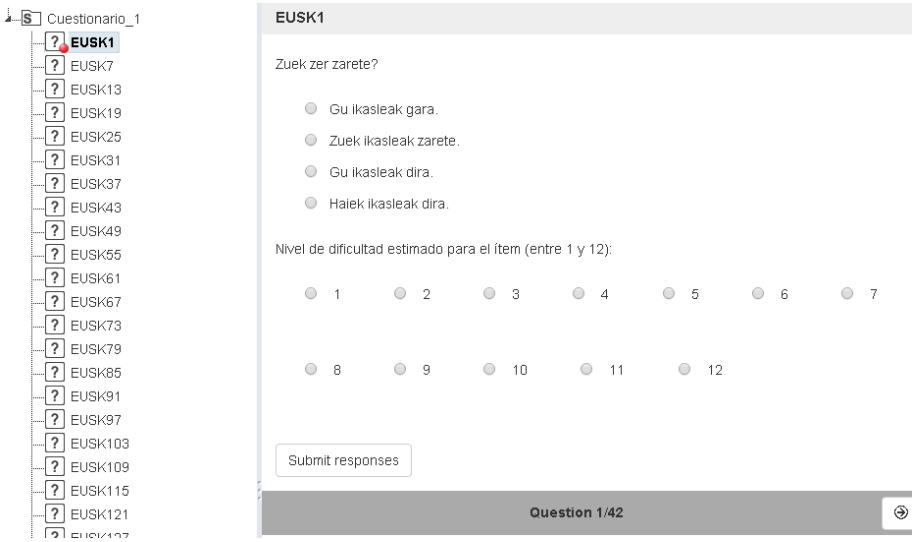


Figura 18 – Presentación de un cuestionario para calibración vía expertos.

En segundo lugar, después de crear el material educativo, ADMINQ Factory da de alta a los participantes, los matricula en el curso y les asigna los distintos cuestionarios. Para cada petición de calibración, crea en Moodle las cuentas de acceso para el responsable y los expertos que se van a ver involucrados en esa calibración. El responsable será matriculado con el rol de profesor en dicho curso y cada experto será matriculado con el rol de alumno. La Figura 19 muestra el mismo curso de la Figura 17 en el que se ha matriculado al responsable (el primero, que es el que ha entrado en Moodle) y a 5 expertos. En el curso del ejemplo, ADMINQ Factory asignaría los dos cuestionarios a estos cinco expertos de forma equitativa, esto es, el Cuestionario 1 a los tres primeros expertos y el Cuestionario 2 a los dos restantes.

Figura 19 – Página de Moodle correspondiente a los participantes del curso.

IX 3.3 CALLIE-MOODLE: Curso y cuentas

CALLIE-MOODLE se encarga de las tareas de administración de los cuestionarios a los expertos. Es un sistema basado en la plataforma Moodle integrada en el sistema de calibración para realizar la parte de administración de los cuestionarios. En el sistema CALLIE-MOODLE se ha ampliado la funcionalidad original de Moodle con servicios web que permiten al módulo ADMINQ Factory de CALLIE-EXPERT colaborar con CALLIE-MOODLE y desplegar los cursos y usuarios en dicho módulo. La plataforma se encuentra por defecto en el mismo servidor que CALLIE, está preparada para acomodarse a las necesidades de administración a expertos y tiene *activados ciertos servicios y opciones* de Moodle útiles para la administración, como por ejemplo, el servicio de mensajería, los servicios web a utilizar, los idiomas más adecuados para los usuarios potenciales o los principales métodos de exportación de informes. Acoge todos los artefactos software generados mediante ADMINQ Factory, es decir, cada curso correspondiente a cada calibración junto con las cuentas de acceso de sus participantes. Con cada curso se da servicio a dos tipos de usuarios, los expertos que van a responder a los ítems y el responsable del experimento.

Respecto a su comportamiento, CALLIE-MOODLE permite adaptar la administración a nuevos requisitos que puedan surgir a partir de la configuración inicial creada por ADMINQ Factory, en función de las distintas opciones que determine el

responsable. Los datos relevantes para la calibración pasarán desde CALLIE-MOODLE al workflow CA mediante PRO-C, a través de un modelo conforme a MMAP.

El responsable de la calibración disfruta de plenos privilegios sobre su curso lo que le permite realizar cualquier modificación a todos los niveles, tanto en la configuración inicial del curso y cuestionarios, como en los expertos matriculados. Por ejemplo, el responsable puede cambiar todas las especificaciones iniciales dadas por ADMINQ Factory: el formato, el idioma, la disponibilidad del curso (visibilidad), los nombres tanto del curso como de los cuestionarios, y/o la asignación de cuestionarios a expertos. También puede modificar las opciones de cumplimentación de cuestionarios: obligar a seguir un cierto orden, no permitir varios intentos, etc. Asímismo, puede establecer/modificar los plazos de disponibilidad del curso, incluir documentos en el curso (como instrucciones de cumplimentación, ejemplos, etc.). A nivel de los expertos involucrados puede añadir otros datos personales a pedir y entrar en contacto con ellos enviándoles mensajes.

Por otro lado, durante el transcurso de la administración, CALLIE-MOODLE guarda automáticamente, en su propia base de datos, todos los datos relevantes sobre el curso y sobre la administración, lo que incluye toda la información aportada por los expertos, tanto sus datos personales como las respuestas dadas. También permite recuperar automáticamente los *datos* almacenados y generar distintos *informes* clasificados por expertos, por ítems, por intentos, por cuestionarios, etc. Gracias a esto el responsable puede, en cualquier momento hasta el cierre, acceder a su curso y comprobar tanto el progreso de las administraciones como la actividad de los expertos. También podrá tomar ciertas decisiones respecto a la evolución del experimento, como la invalidación de una administración realizada a un experto concreto; suspender y reanudar el acceso a uno o varios cuestionarios; incluso puede dar por finalizado el curso a la vista de la evolución del mismo. Al tener la posibilidad de *ver múltiples informes* de respuestas detalladas, estadísticas, tiempos invertidos, actividad de los expertos, etc. puede *realizar cualquier tipo de análisis manual* sobre las entradas almacenadas o sobre cada cuestionario cumplimentado y reflejarlo en la plataforma utilizando las opciones de calificación manual o de eliminación de intentos y de matriculación/desmatriculación de expertos. Así, puede eliminar tanto ítems como expertos. Finalmente, cuando se genera un curso en Moodle mediante la importación de un paquete IMS el banco de ítems se duplica en esta plataforma, lo que permite al responsable *cambiar el diseño del experimento*, modificando cuestionarios existentes y administrándoselos a otros expertos si ve que dispone de pocos datos o que algunos no cumplen con lo especificado.

IX 3.4 CALLIE-EXPERT y WWF: WF Factory y Workflow CA

WF Factory es el componente encargado de generar un workflow que realiza el análisis de datos y calibración en base a un modelo conforme a MMANCA. Este workflow, denominado Workflow CA, se ejecuta mediante un motor Windows Workflow Foundation (WWF, el motor de Workflows de Microsoft) y produce resultados conformes al MMRE. Este componente dispone de una *biblioteca de análisis y cálculos* predefinidos que el responsable podrá seleccionar a voluntad.

Respecto a su comportamiento, el **workflow CA** consta de dos componentes que se ejecutan de manera consecutiva (Figura 20). El primero – *AdmAnalyzer* – realiza los

análisis de administraciones que no son exclusivos de una calibración vía expertos a partir de los datos de entrada conformes al MMAP. El segundo – *AnalyzerCalibrator* – lleva a cabo los análisis de fiabilidad propios de una calibración vía expertos y la estimación de la dificultad de los ítems no descartados. Como resultado, se obtienen los datos detallados de los análisis y estudios realizados y los datos y resultados finales de los cálculos de calibración.

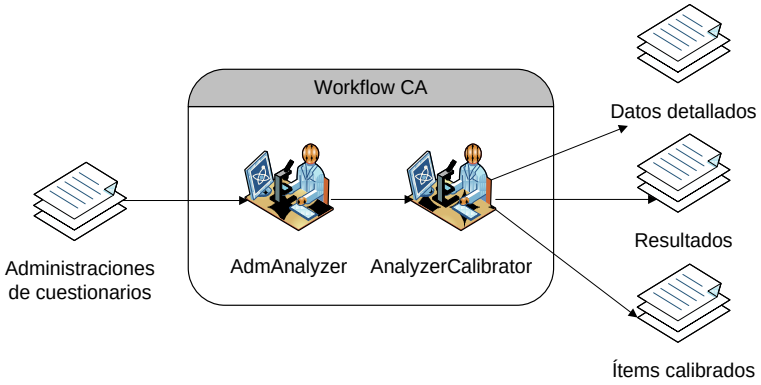


Figura 20 – Arquitectura del workflow CA.

IX 3.5 CALLIE-EXPERT: PRO-C

PRO-C interactúa con el responsable, mediante su interfaz CALLIE-PRO, de la que se hablará en la sección X 2, y le permite controlar la ejecución del proceso asociado a cada una de sus peticiones de calibración ya confirmadas.

PRO-C permite al responsable transferir el modelo de las aportaciones de los expertos al sistema WWF. Además, permite controlar y monitorizar las tareas que el sistema WWF realiza con dicho modelo. Para ello, el módulo dispone de las siguientes funcionalidades:

(1) *Recoger (más) datos.* Recoge los datos conforme al MMAP desde el susbsistema de administración utilizado, esto es, CALLIE-MOODLE u otro. Sirve para recoger los datos una primera vez y también para recoger más administraciones si la muestra es insuficiente después de una depuración. Si el subsistema es CALLIE-MOODLE, PRO-C recoge las *administraciones de cuestionarios* desde Moodle y las prepara para su análisis a instancias del responsable de la calibración. Partiendo del modelo de datos conforme al MMCU recoge, para el curso asociado con esa calibración, los datos de las administraciones *terminadas* correspondientes a los expertos del curso y todos los ítems *intentados*, y los transforma al formato de datos que necesita el workflow CA conforme al MMAP. Si el subsistema es otro, PRO-C recoge los datos conforme al MMAP directamente desde una hoja Excel.

(2) *Depurar y Calibrar*. Invoca la ejecución del Workflow CA con los parámetros especificados para el MMANCA en el modelo de calibración. Ofrece un modelo de resultados conformes al MMRE.

(3) *Aceptar/Rechazar Resultados.* Ambas concluyen la fase de calibración. *Aceptar resultados* terminará la calibración con éxito creando el banco de ítems calibrado con las estimaciones de dificultad que pudieron calcularse, que pasan de

forma definitiva a la base de datos de CALLIE-EXPERT. Por su parte, *Rechazar resultados* también la terminará pero sin almacenar los resultados.

(4) *Abortar*. Aborta el proceso de calibración para la petición indicada, en cualquier momento entre su aceptación y su finalización.

Este módulo dispone además de una funcionalidad adicional para obtener resultados alternativos conformes al MMRE, aplicando distintos cribados a la misma muestra de datos. Esta funcionalidad, denominada *Cambiar Filtros y Simular*, solicita al responsable nuevos parámetros conformes al MMANCA y después invoca con ellos la ejecución del workflow CA. Esta opción permite, por ejemplo, saber si sería necesario recoger más datos o si sería mejor establecer otros filtros menos restrictivos.

Cabe destacar que la interfaz inicial para la funcionalidad *Recoger (más) datos desde Excel* era muy rudimentaria, ya que solo ofrecía ayuda mediante un texto en pantalla, no explicaba los distintos pasos a dar y obligaba al usuario a construir desde cero el fichero Excel, en un formato de columnas preestablecido, una ubicación concreta y con un nombre determinado. Así, una vez concluidas las primeras pruebas con usuarios y debido a los malos resultados obtenidos usando esta interfaz inicial, se mejoró el componente de recogida de datos desde Excel. Concretamente la autora de esta memoria implementó una interfaz mejorada, que ahora aglutina los distintos pasos a dar y además incluye la generación automática de una plantilla con el formato adecuado para la copia y subida de los datos recopilados a los expertos desde otras fuentes diferentes a CALLIE-MOODLE (como se puede ver en la página web que se muestra en la Figura 44 de la memoria).

IX 4 El módulo EXT

El módulo EXT es un componente adicional de CALLIE que complementa ESKARI y PRO-C. El objeto de EXT es aglutinar distintas opciones de comunicación de CALLIE con sistemas externos, como Moodle u otras herramientas, y ponerlas a disposición del responsable a través de su interfaz CALLIE-EXT. Las funcionalidades que ofrece el módulo son:

(1) *Importación/exportación de peticiones de calibración* en formato XML conformes al modelo MMCA.

(2) *Recogida de datos mejorada desde hojas de Excel*.

(3) *Gestión de los expertos participantes* en una determinada calibración con CALLIE-MOODLE mediante la creación, matriculación y eliminación de expertos en sus cursos a través de servicios web de Moodle.

Este módulo ha sido el resultado de un trabajo de fin de grado dirigido por la autora de esta memoria. En su realización se han aprovechado ciertas utilidades ya existentes en CALLIE-EXPERT, tales como la recogida mejorada de datos desde Excel, y se han añadido otras nuevas. Todos los detalles sobre EXT se pueden encontrar en Irastorza (2014).

X Interfaz de CALLIE-EXPERT

La interfaz de CALLIE surge tras el estudio de la comunicación utilizando un modelo de género del sistema de aplicación de test adaptativos informatizados (López-Cuadrado, Armendariz, Latapy y Lopistéguy, 2008) y las lecciones aprendidas en un estudio empírico sobre la usabilidad de aplicaciones software (Presedo, Dolado y Aguirregoitia, 2010) empleado para estudiar la herramienta MetroMap (Aguirregoitia, Dolado y Presedo, 2008b; Aguirregoitia, Dolado y Presedo, 2010a) y otras herramientas que siguen metáforas de visualización orientadas al control y seguimiento de tareas (Aguirregoitia, Dolado y Presedo, 2010b).

El resultado es una interfaz *flexible* porque, aunque se sugieren una serie de decisiones a tomar de manera secuencial, el usuario puede alterar el orden indicado según sus preferencias o necesidades. Es *informal* ya que el lenguaje que utiliza se centra en conceptos de los sistemas de aprendizaje, minimizando el uso de términos técnicos de psicometría que se hacen necesarios para acometer la calibración. También es *sencilla*, puesto que la presentación de las decisiones a tomar se hace siguiendo la metáfora del asistente, en el que la tarea a ejecutar se descompone en una serie de pasos a realizar de manera secuencial y, para incidir en la prevención de errores, cuando existen varias opciones se presentan listas cerradas con esos valores únicamente. Asímismo, dispone de una opción en forma de resumen mediante la que se puede obtener información general sobre las decisiones tomadas para la calibración. Por último, a la hora de hacer las especificaciones CALLIE-EXPERT ofrece guía adaptativa porque el sistema adapta las opciones que ofrece a lo ya decidido en pasos anteriores. El asistente además vigila que se tomen todas las decisiones. La adaptatividad detecta si una decisión afecta a las ya tomadas y, en tal caso, advierte de los efectos que suponen las mismas. Por otro lado, el sistema siempre recomienda una opción válida (según el criterio del sistema) para cada decisión que se debe tomar. Si el usuario modifica esta recomendación, CALLIE comprueba el cumplimiento de las restricciones que se deben seguir para ese paso, avisando de los posibles problemas que surjan. Además, proporciona al usuario un tipo de ayuda que se genera automáticamente cuando el responsable de la calibración debe establecer valores y realizar cálculos. Como ejemplo de esta adaptatividad, el sistema CALLIE-EXPERT cambia automáticamente el número medio de ítems por cuestionario si el usuario modifica la cantidad de cuestionarios a diseñar o avisa si con las decisiones tomadas no va a ser posible conseguir las 7 valoraciones por ítem recomendadas por la herramienta. Todas estas recomendaciones, avisos y cálculos automáticos surgen de los valores por defecto que la herramienta mantiene asociados a sus metamodelos. Así, en caso de problemas, siempre se puede restablecer la opción válida sugerida por el sistema. Se pueden encontrar ejemplos de todas estas características en las figuras que ilustran los siguientes apartados.

La interfaz ofrece varios tipos de ayuda: contextual, con colores y asistente adaptativo (o guía adaptativa). Ofrece ayuda contextual sobre cualquier objeto de la pantalla al situar el ratón sobre el elemento en cuestión. La herramienta CALLIE utiliza

un código de colores (rojo, naranja y verde), mediante el que avisa si hay algún problema asociado a algún valor. Este código de colores se acompaña con los avisos textuales oportunos para que el usuario visualice rápidamente las incidencias que le ocurran durante la introducción de datos. En la interfaz también aparecen valores sombreados que representan valores predefinidos no aplicables al caso actual o no modificables (fondo gris) o, si sí son modificables, que no se pueden modificar en ese paso, generalmente porque se han decidido con anterioridad (fondo azul). En una determinada pantalla los valores calculados por defecto por CALLIE-EXPERT que el responsable podrá cambiar aparecen con fondo blanco.

Las páginas Web que genera CALLIE siguen el mismo estilo y se han diseñado e implementado de forma que siempre aparezca la página completa en la pantalla del ordenador (i.e. sin necesidad de hacer desplazamientos ni verticales ni laterales).

Al igual que ocurre con la arquitectura del sistema, la interfaz de CALLIE posee dos partes bien diferenciadas, que forman parte de los componentes de la arquitectura con el mismo sufijo: CALLIE-ESKARI y CALLIE-PRO. Estas dos partes son dos submenús de la aplicación principal, CALLIE, que permiten al responsable, por un lado, diseñar una o varias calibraciones y, por otro, controlar su ejecución una vez creados los artefactos correspondientes.

A continuación, se describen las características y posibilidades de estas dos interfaces de CALLIE. Este capítulo se centra en la interfaz de CALLIE-EXPERT mientras que la interfaz de CALLIE-TRI está descrita en Armendariz (2014).

X 1 Interfaz de ESKARI: CALLIE-ESKARI

CALLIE-ESKARI guía al responsable en la introducción de todos los datos necesarios para diseñar su calibración, utilizando un menú de cuatro pasos como elemento principal. Estos cuatro pasos tienen sentido porque CALLIE-EXPERT se integra con CALLIE-TRI de modo que los dos primeros pasos son comunes a ambos tipos de calibración, son los primeros que se dan y una vez dados ya no se pueden cambiar. El usuario puede rectificar las decisiones tomadas en los otros dos pasos en todo momento hasta que las acepte, enviando la petición al sistema. Mientras no se completen todos los pasos, el sistema no permite el envío de la petición, lo que asegura que el sistema no la registra antes de que se hayan establecido todas las especificaciones necesarias para generar el modelo de calibración correspondiente.

Cada paso posee sus correspondientes pantallas y permite recoger unos datos muy concretos. Los cuatro pasos son los siguientes: el banco de ítems a calibrar (Paso 1), el tipo de calibración que se desea efectuar (Paso 2), los cuestionarios para administrar a los expertos (Paso 3) y otras decisiones de administración, análisis y calibración, tanto sobre los datos a pedir por ítem como sobre el tratamiento automático que el sistema dará a los datos recogidos de los expertos (Paso 4).

El mapa con los pasos del asistente de CALLIE-EXPERT para CALLIE-ESKARI se muestra en la Figura 21 y se detalla a continuación.

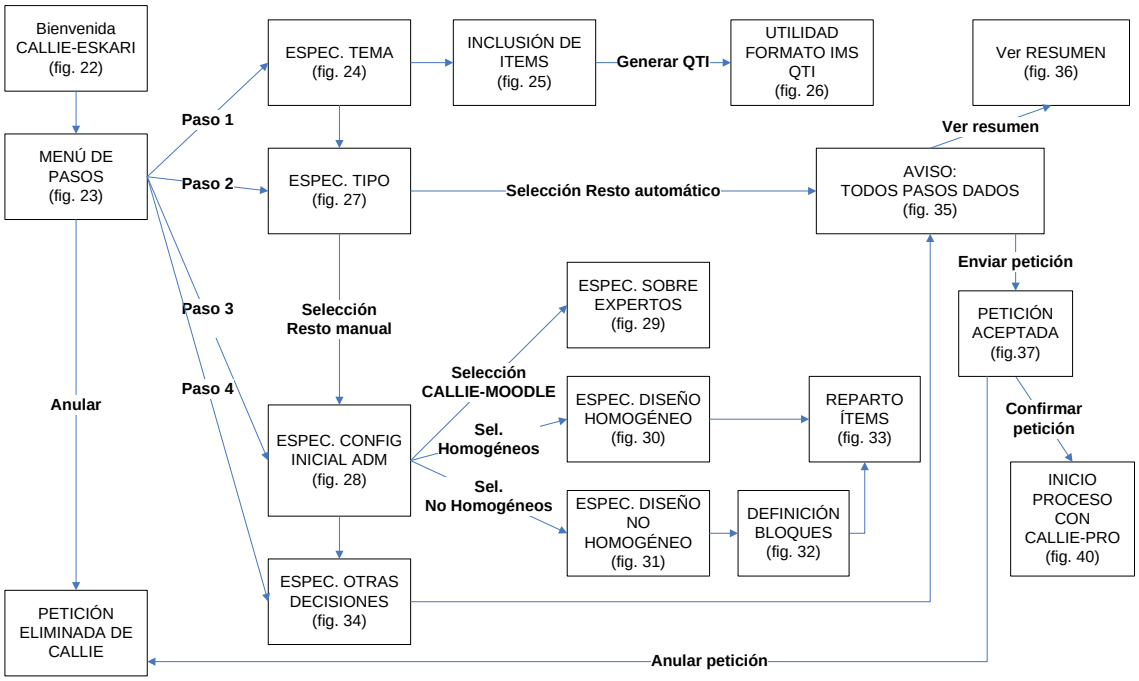


Figura 21 – Mapa con los pasos del asistente para CALLIE-ESKARI.

CALLIE-ESKARI comienza con la página de bienvenida (Figura 22) en la que además se ofrece al responsable la posibilidad de simular todo el proceso utilizando el banco de pruebas EUSK con un número de ítems a elegir (desde 1 a 252), lo que permite al usuario probar el sistema antes de introducir sus propios ítems.

Figura 22 – CALLIE-ESKARI: Página de bienvenida.

Cuando el usuario pulsa *Continuar* en la página de Bienvenida de CALLIE-ESKARI, es redirigido a la página "CALLIE-ESKARI: Pasos y estado de la petición de calibración" (Figura 23) que consta del menú con cuatro pasos en el que se indica al

responsable qué pasos se han dado ya y cuál sería el siguiente paso recomendado. CALLIE recomienda dar estos cuatro pasos en el orden de aparición. También hay que destacar que al terminar cualquier paso del proceso, el sistema ofrece al usuario la opción de *anular todo y volver a comenzar*. En caso de anulación, la petición desaparece del sistema.

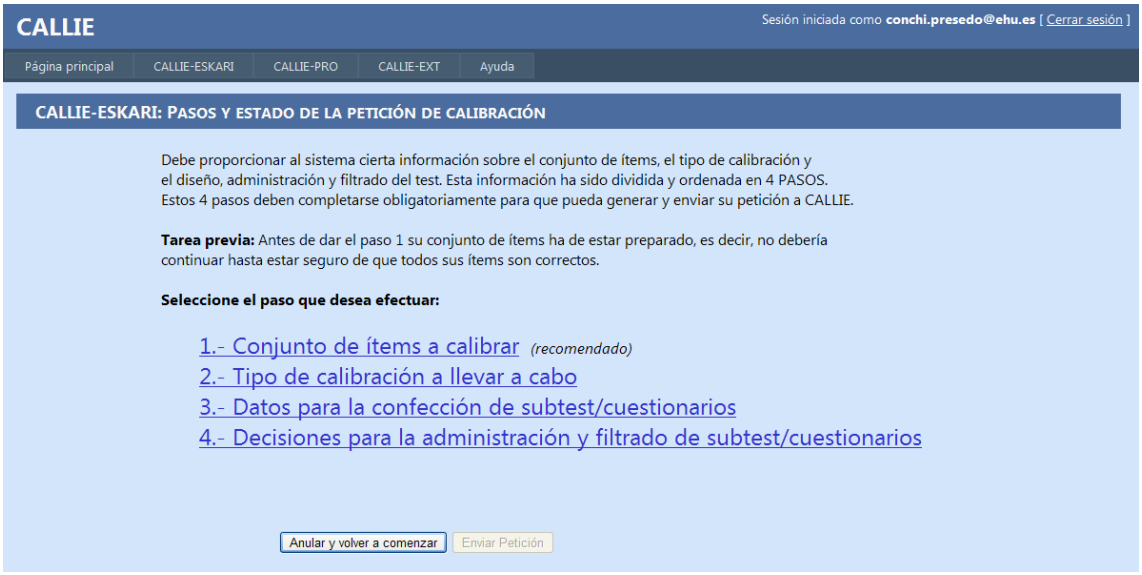


Figura 23 – CALLIE-ESKARI: Página de Menú de pasos.

La selección del **paso 1** – Conjunto de ítems a calibrar – del menú llama a tres páginas web que posibilitan la especificación del tema y los ítems siguiendo el modelo de datos conforme al MMCU. La primera página (Figura 24) permite especificar un nuevo tema con sus propias categorías o seleccionar alguno ya existente. Como ejemplo, en la figura se crea un nuevo tema – IRALE – con sus ítems en formato IMS QTI perfil CC, 7 categorías y 3 niveles previos.

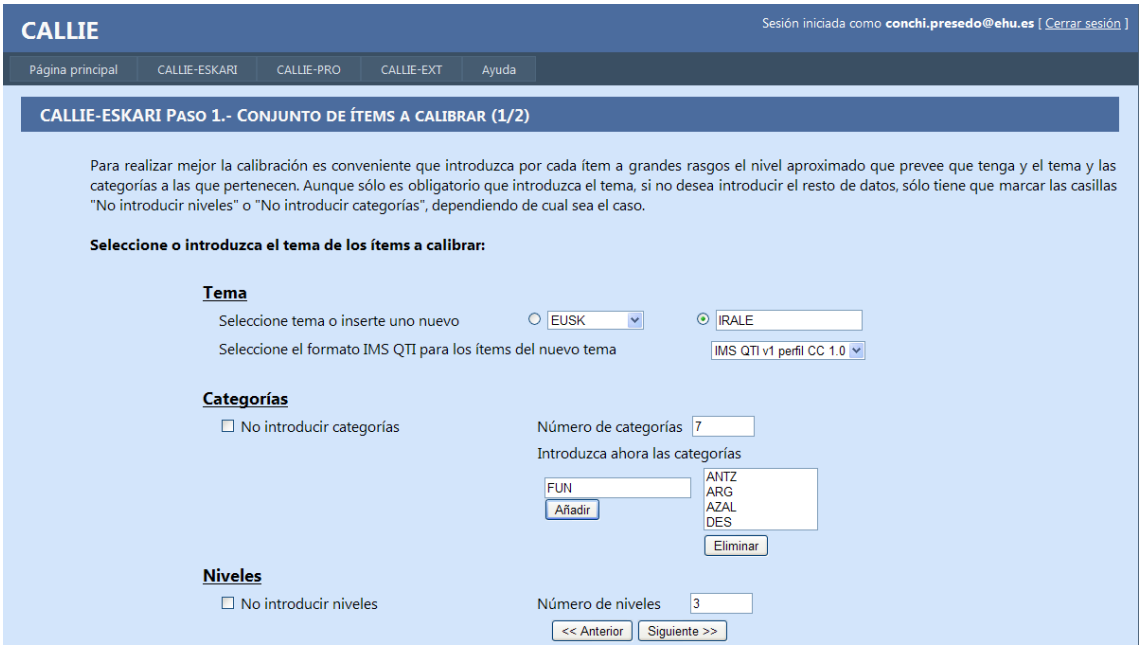


Figura 24 – CALLIE-ESKARI: Página de especificación del tema a calibrar.

La segunda página (Figura 25) permite visualizar la lista actual de ítems en el banco y añadir nuevos ítems al banco seleccionado. En el ejemplo, ya se han introducido 132 ítems en el tema IRALE, resultando el banco que se visualiza en la figura. La caja de texto y el botón *Añadir* permiten importar ítems que han de estar en formato IMS QTI, y el botón de *Ayuda* muestra ayuda sobre ese formato. El botón *Finalizar* termina la introducción de ítems y vuelve al menú.

Figura 25 – CALLIE-ESKARI: Página de inclusión de ítems.

Por su parte, el botón *Generar QTI* de la Figura 25 redirige a la tercera página web de este paso (Figura 26): una utilidad que permite la adición automática de ítems de selección múltiple en el formato IMS QTI elegido para el banco.

Figura 26 – CALLIE-ESKARI: Página de introducción de ítems de selección múltiple en IMS QTI.

La selección del **paso 2** – Tipo de calibración a llevar a cabo – del menú llama a una única página web (Figura 27) que posibilita la especificación del método de calibración. Además permite especificar si el resto de pasos (3 y 4) los decidirá el sistema automáticamente, utilizando sus valores por defecto, o los introducirá el responsable manualmente.

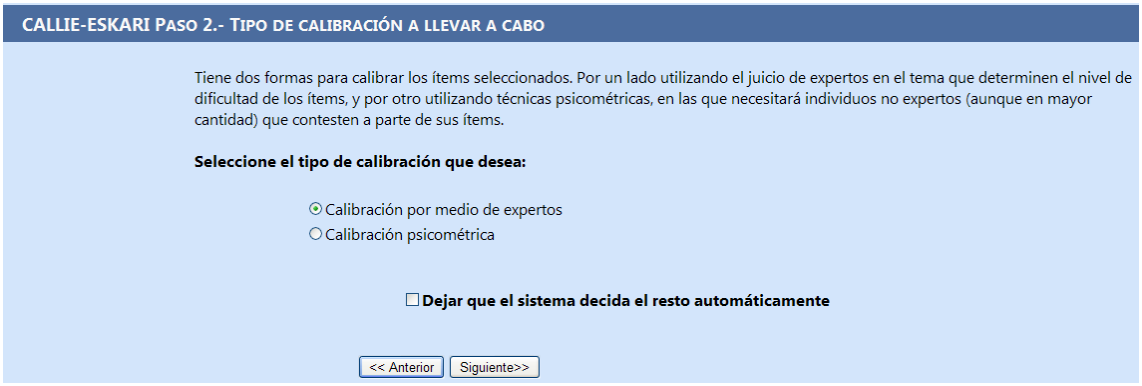


Figura 27 – CALLIE-ESKARI: Página de especificación del tipo de calibración.

La selección del **paso 3** – Datos para la confección de subtest/cuestionarios – del menú llama a seis páginas web en las que se determinan los datos necesarios del modelo de datos conforme al MMCA y se reparten los ítems en cuestionarios siguiendo el modelo de datos conforme al MMCU. La primera página permite especificar la configuración inicial para la administración (Figura 28).

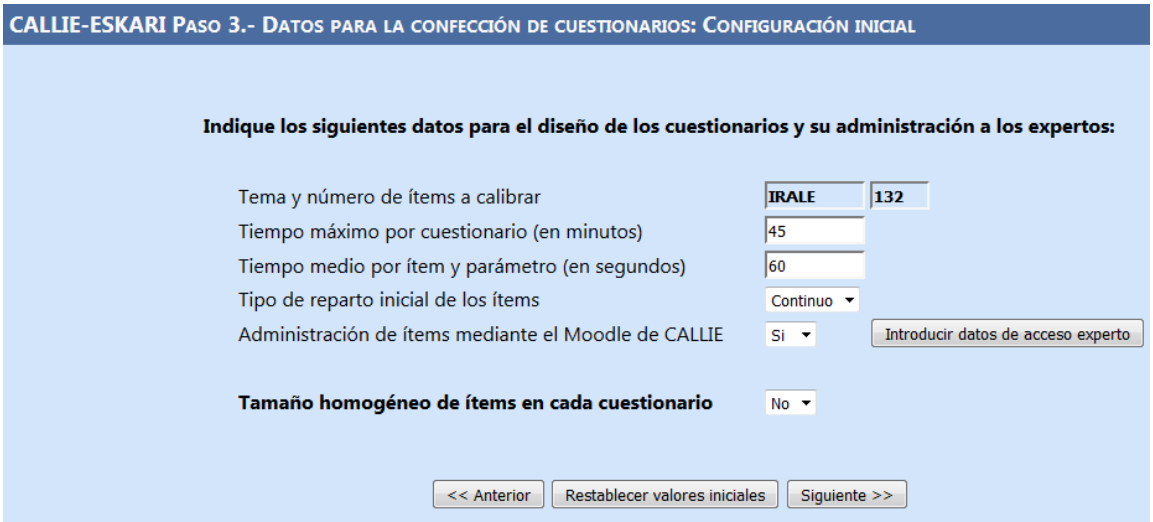


Figura 28 – CALLIE-ESKARI: Página de configuración inicial para la administración.

Además, en caso de administración con CALLIE-MOODLE, el botón *Introducir datos de acceso experto* lleva a la pantalla que permite especificar los datos relevantes sobre los expertos (Figura 29).

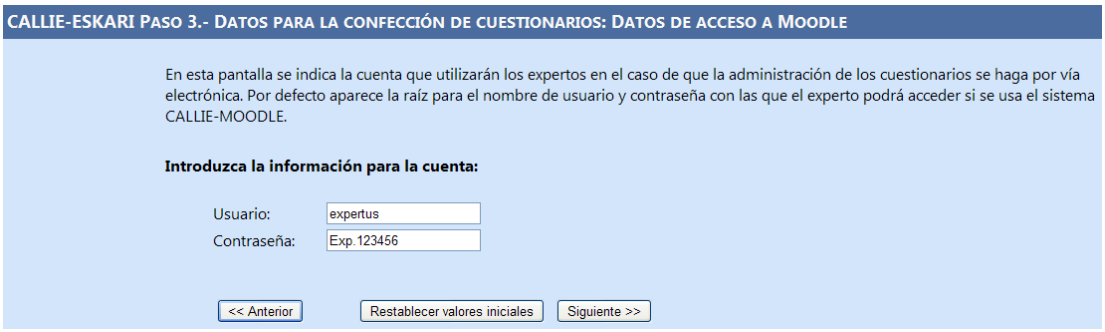


Figura 29 – CALLIE-ESKARI: Página de especificación de datos sobre los expertos.

Continuando con el paso 3, se muestran distintas páginas que dependen de si se ha elegido un tamaño homogéneo o no para los cuestionarios en la Figura 28.

La configuración de **cuestionarios de tamaño homogéneo** consta de una única pantalla en la que se indicarán los parámetros específicos para su diseño (Figura 30). En el ejemplo se puede ver una pantalla de CALLIE-EXPERT en la que se ha indicado que hay 30 expertos disponibles, lo que da un número de opiniones por ítem de 5. Estos datos han activado en naranja los datos que han hecho saltar esas alertas junto con los avisos que indican que se incumplen el número de expertos necesarios para los 6 cuestionarios (42) y el mínimo de opiniones recomendadas a recabar (7).

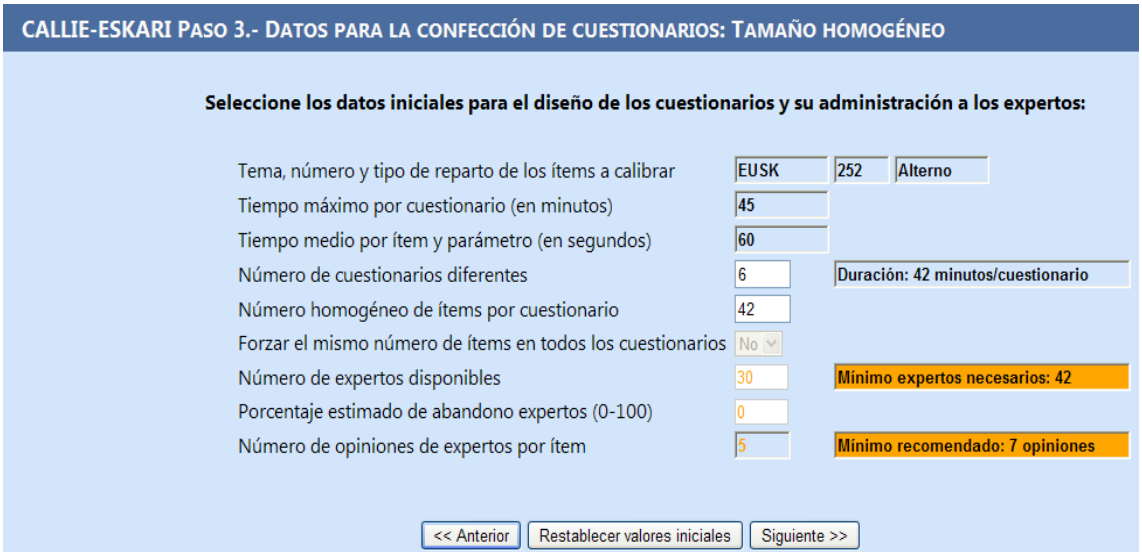


Figura 30 – CALLIE-ESKARI: Página de especificación de diseño homogéneo.

Si se han elegido cuestionarios de tamaño **no homogéneo** se indican los parámetros específicos para diseñar estos cuestionarios mediante dos pantallas sucesivas. En la primera (Figura 31) se especifica el diseño no homogéneo, explicitando el número de bloques a diseñar y la cantidad de expertos disponibles.

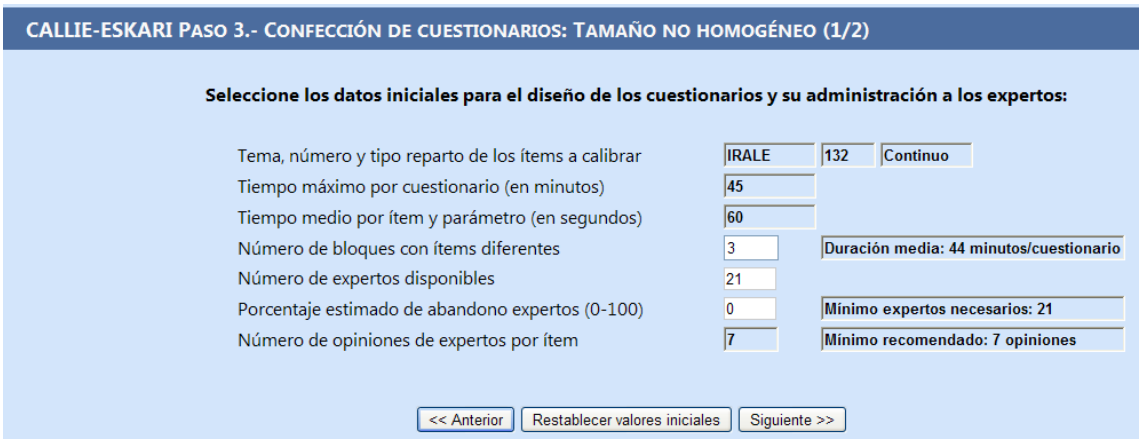


Figura 31 – CALLIE-ESKARI: Página de especificación de diseño no homogéneo.

En la segunda página (Figura 32) se definen los bloques, esto es, se indica la composición numérica en ítems de los distintos bloques y cuestionarios.

Figura 32 – CALLIE-ESKARI: Página de definición de bloques.

Durante la cumplimentación de estas tres últimas pantallas, el sistema realiza cálculos para detectar anomalías en el diseño de los cuestionarios, y que pueden dar lugar a distintas acciones si existe alguna. Por un lado, detecta *situaciones de alerta o de aviso* que se corresponden con las violaciones en las recomendaciones que deberían cumplirse para respetar los tiempos estimados y conseguir el mínimo de juicios por ítem. Aunque no se impide que el responsable continúe con el diseño, el sistema le avisa mediante un código naranja. Concretamente, detecta si la duración de los cuestionarios con los valores actuales supera la duración máxima establecida (*Alerta 1: Duración del cuestionario recomendada*), si el número de expertos disponibles es menor que el recomendado para obtener el número mínimo de opiniones por ítem (*Alerta 2. Mínimo de expertos recomendados*), y si el número medio de valoraciones por ítem actual es menor que el mínimo recomendado (*Alerta 3. Mínimo de valoraciones recomendadas*).

Por otro lado, también revisa *restricciones insalvables o de error* que aparecerán en código rojo y deshabilitan el botón *Siguiente>>* en la pantalla correspondiente, lo que impide al responsable continuar hasta su corrección. Concretamente, verifica que todos los datos introducidos estén en el rango adecuado, que el total de ítems de los bloques no coincida con la cantidad de ítems del banco, que todos los ítems a calibrar formen parte de alguno de los cuestionarios/bloques a administrar y que cada uno de estos cuestionarios/bloques contenga la cantidad de ítems previamente especificada.

Al pulsar *Siguiente>>* en la última página de cada caso (Figura 30 o Figura 32), se completa el paso 3 con una sexta pantalla de **reparto de los ítems** (Figura 33), en la que el responsable permacerá hasta haber asignado todos los ítems. En el ejemplo de la figura se indica que el cuestionario/bloque actual es el número 1 que debe constar de 42 ítems distintos (decidido anteriormente), los ítems asignados a este cuestionario/bloque y sus posiciones aparecen en la parte inferior de la pantalla bajo la etiqueta *Items del cuestionario/bloque nº1* en color blanco y azul. En el panel *Estado de cada cuestionario/bloque* se indica que existen 6 cuestionarios/bloques a rellenar, que algunos tienen ya todos los ítems asignados (en verde) y que faltan por asignar (en rojo) 3 y 1 ítems en los cuestionarios/bloques 2 y 3 respectivamente. En la parte central izquierda se puede ver la lista de los 4 ítems no asignados todavía.

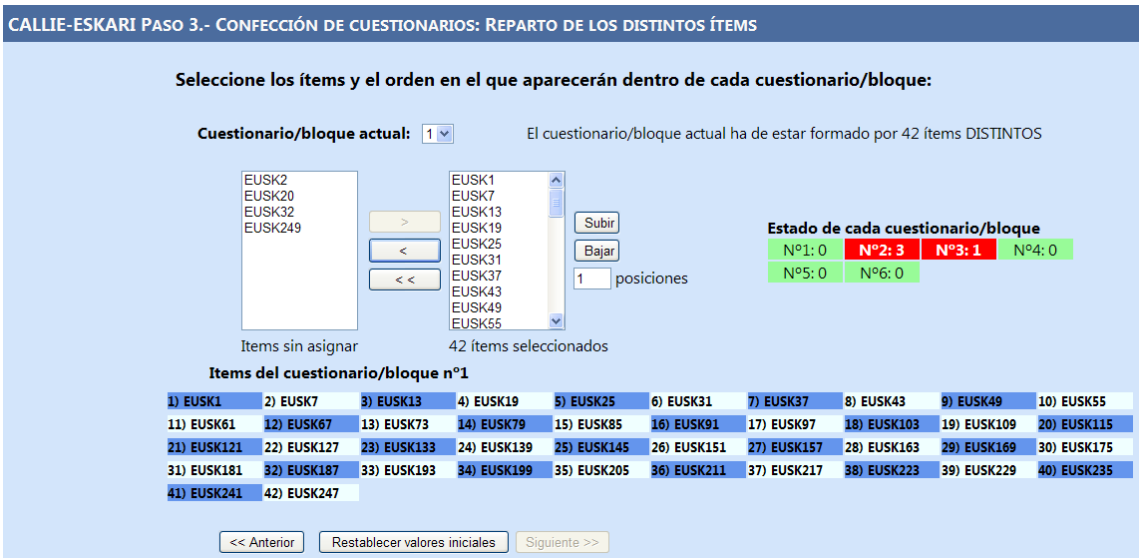


Figura 33 – CALLIE-ESKARI: Página de reparto de ítems.

La selección del **paso 4** – Decisiones para la administración y filtrado de subtest/cuestionarios – del menú llama a una única página web (Figura 34) que posibilita la especificación de los datos del modelo de datos conforme al MMCA que delimitan la actuación del sistema de administración y los parámetros del modelo de datos conforme al MMANCA. Concretamente, se determinan los datos necesarios para diseñar el formato final de los cuestionarios a administrar y el filtrado posterior de los ítems o administraciones recogidas. En la columna izquierda de la página se indica el estadístico a aplicar para estimar la dificultad de cada ítem así como los datos a pedir por ítem y sus características, mientras que en la columna derecha se indican los parámetros para el análisis sobre los datos recogidos, que dependen de la biblioteca de análisis y cálculos y que estarán condicionados por las decisiones tomadas en la parte izquierda.

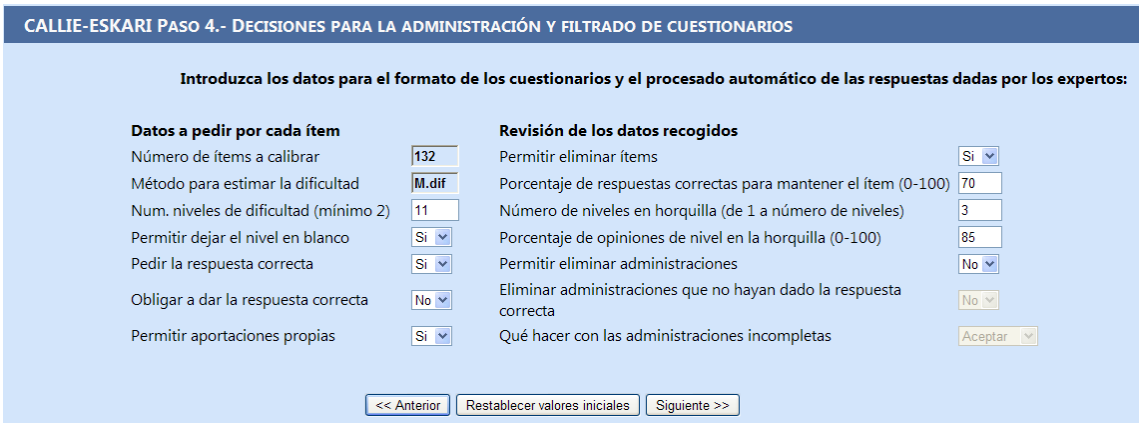


Figura 34 – CALLIE-ESKARI: Página de otras decisiones de administración y filtrado.

Cada vez que un paso se completa, CALLIE-ESKARI vuelve al menú y ese paso se muestra como dado. Una vez cumplimentados los cuatro pasos por primera vez (Figura 35), se activan el enlace *Ver resumen petición* y el botón *Enviar petición*.

Figura 35 – CALLIE-ESKARI: Menú de pasos con los cuatro pasos dados.

Antes de enviar la petición, el usuario puede repasar las elecciones efectuadas mediante la opción *Ver resumen petición* que consta de una página web en la que se muestran todas las decisiones relevantes tomadas y el código que identifica la petición, de modo similar al indicado en la Figura 36. Existe la posibilidad de *Imprimir* este resumen.

La opción *Enviar Petición* realiza la tarea del mismo nombre, que genera el modelo de calibración en XML (ver ejemplo en el Anexo A1), y redirige al usuario a una nueva página (Figura 37) para que confirme que desea continuar. En esta pantalla de confirmación se avisa al responsable de que su petición ha sido aceptada, y se le

indican los siguientes pasos a dar. Si el sistema de administración elegido fue CALLIE-MOODLE, se le indican los datos del curso y las distintas cuentas que se han creado para los participantes. Con la opción *Confirmar Petición* CALLIE-ESKARI realiza la tarea Iniciar Proceso y redirige al usuario a CALLIE-PRO. La opción *Anular Petición* elimina la petición del sistema.

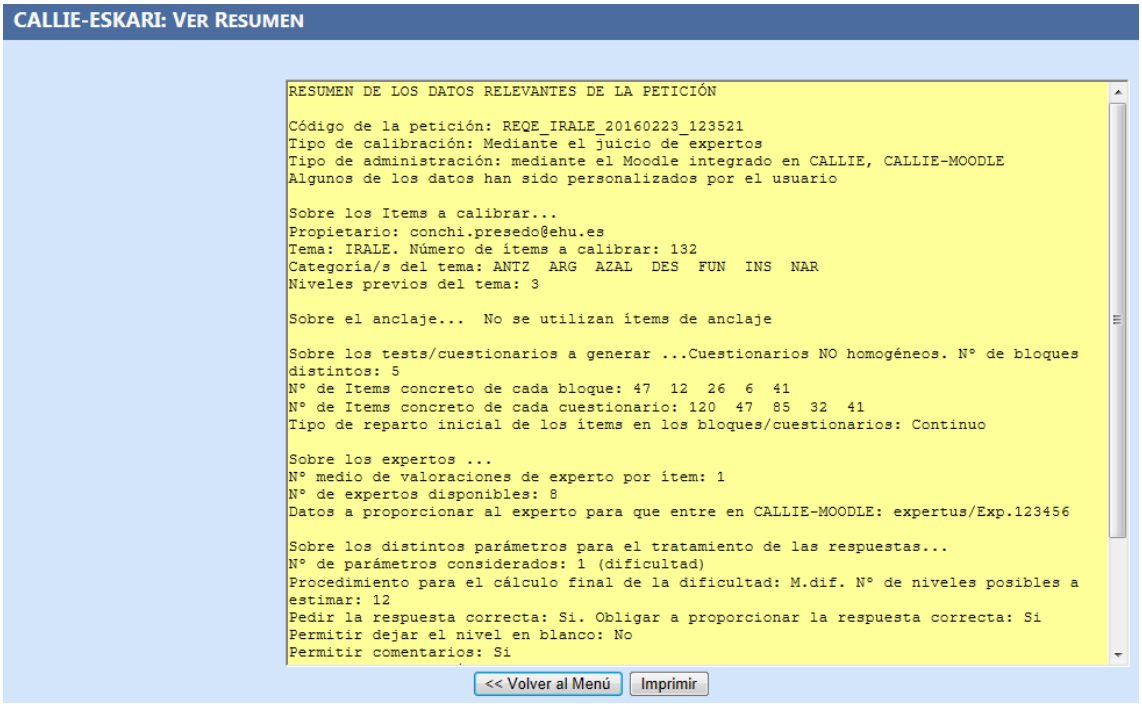


Figura 36 – CALLIE-ESKARI: Página ver resumen.

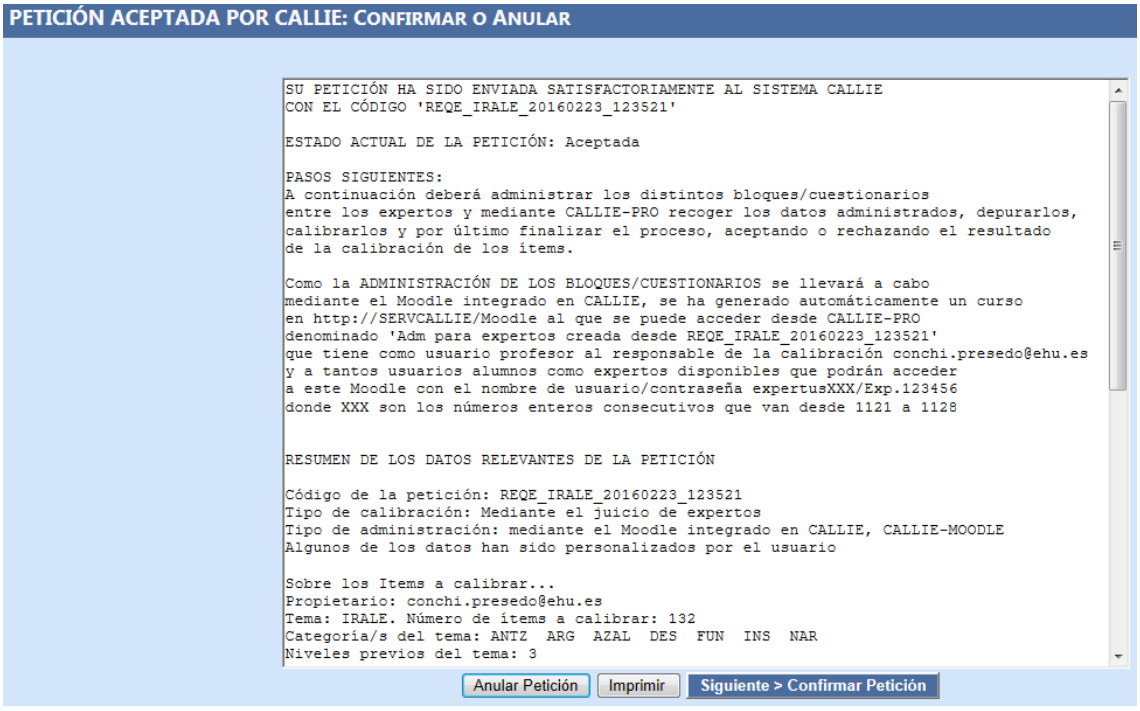


Figura 37 – CALLIE-ESKARI: Página de petición aceptada.

X 2 Interfaz de PRO-C: CALLIE-PRO

La interfaz de PRO-C guía al responsable por el proceso de calibración que él mismo ha diseñado en su petición, utilizando tres páginas web principales y sucesivas: (1) Procesar mis peticiones de calibración, (2) Ver resultado de la calibración y (3) Opciones de finalización. Mediante la primera de ellas (Figura 40), el sistema avisa del estado en que se encuentra cada petición y también avisa de los cambios en ese estado, permite al responsable seleccionar la petición sobre la que desea actuar y la acción a realizar. En la segunda página web (Figura 45) se pueden visualizar los resultados finales obtenidos por la herramienta. Además, si el responsable lo considera oportuno, puede realizar otras actividades sobre la muestra de datos recogidos como: generar un documento con los cálculos detallados realizados por la herramienta, recoger más datos desde el sistema de administración elegido, y cambiar los filtros para simular nuevos resultados de calibración modificando los parámetros de análisis sobre los datos recogidos. Mediante la tercera de estas páginas web (Figura 48), se ofrecen distintas opciones de finalización del proceso para el responsable del tema objeto de la calibración. Mediante otras páginas web y ficheros Excel auxiliares, el sistema informa o ayuda en el propio proceso de calibración.

El mapa con los pasos del asistente de CALLIE-EXPERT para CALLIE-PRO se muestra en la Figura 38.

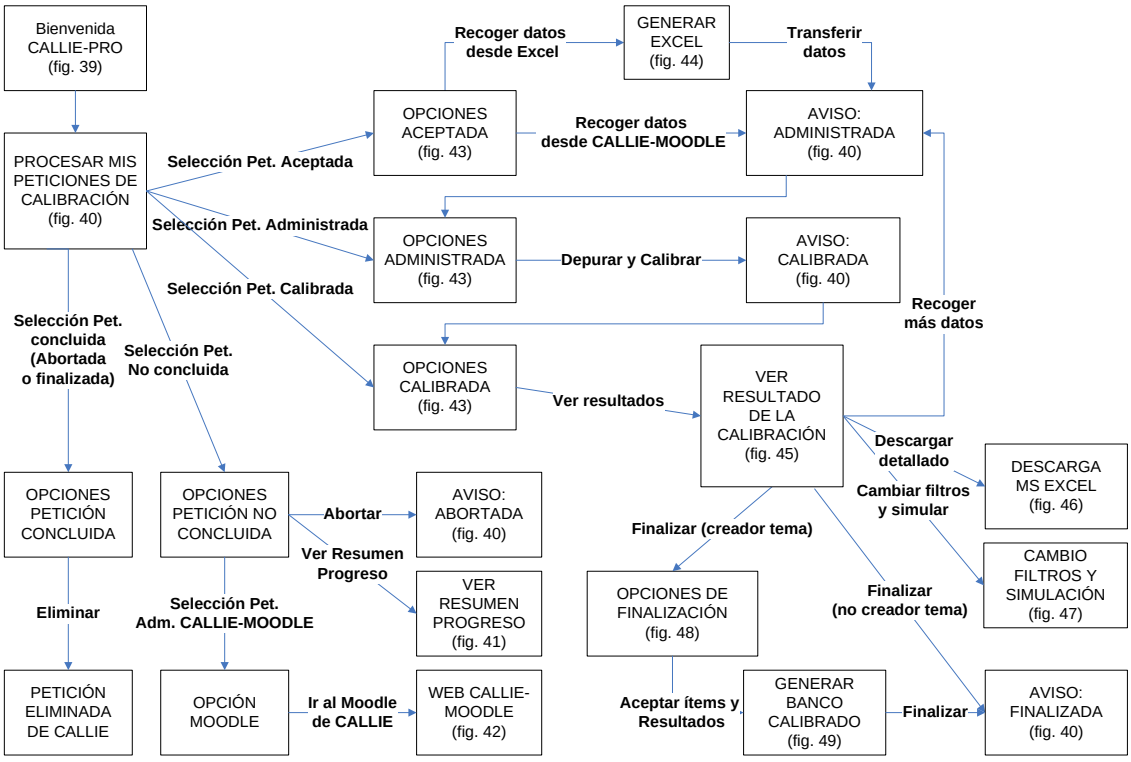


Figura 38 – Mapa con los pasos del asistente para CALLIE-PRO.

CALLIE-PRO comienza con la página de bienvenida al módulo (Figura 39) que le permite gestionar sus peticiones de calibración confirmadas.

Figura 39 – CALLIE-PRO: Página de bienvenida.

Al pulsar *Continuar* aparece la página "Procesar mis peticiones de calibración" que muestra una lista que contiene todas las peticiones de calibración del responsable (Figura 40). El responsable puede ver todas las peticiones que ha confirmado y, en el momento en que establece sobre cuál de ellas quiere operar, CALLIE-PRO le presenta las distintas acciones que puede efectuar. Esta página utiliza colores para indicar la finalización con éxito (verde) o no (rojo) de las calibraciones ya concluidas. En el caso de que el usuario no tuviese peticiones en curso se le indicaría mediante un mensaje en pantalla.

Figura 40 – CALLIE-PRO: Página procesar mis peticiones de calibración.

La *lista de peticiones* que se ve en la Figura 40 muestra varias peticiones en distintos estados que abarcan todo el espectro de opciones del plano. Esta lista contiene los códigos de las peticiones, la fecha y hora a la que se realizaron, el tipo de calibración a realizar, el banco de ítems a calibrar y qué tipo de administración se desea. El código de la petición se muestra en color naranja si se ha superado el plazo máximo predeterminado para la ejecución de todo el proceso (por defecto 180 días) desde que se solicitó. En las dos últimas columnas se indica el estado en el que se encuentra el proceso en este momento, junto con la fecha y hora a la que se alcanzó dicho estado. La lista está ordenada por estos dos campos. Además, cuando PRO-C realiza alguna tarea que implica un cambio en el estado de la calibración (i.e. las tareas asociadas a las opciones Recoger datos, Depurar y calibrar, Abortar y Finalizar), el asistente avisa al responsable resaltando la petición involucrada y actualizando los datos de estos dos últimos campos.

Al *Seleccionar* una petición de esta lista se activan las opciones adecuadas según el estado en el que se encuentre esa petición, y esas opciones se muestran en la última fila de la página "Procesar mis peticiones de calibración". Toda petición concluida, esto es, en estado abortada o finalizada, presenta una única opción *Eliminar*, que hace que todos los datos asociados a la petición desaparezcan del sistema. Cuando una petición no ha concluido, se activan las opciones *Abortar* y *Ver Resumen Progreso* para esa petición. La opción *Abortar* concluye la petición en cualquier momento intermedio del proceso. *Ver Resumen Progreso* (Figura 41) muestra qué es lo que se ha hecho ya y cuál es el siguiente paso a dar si es que lo hay.

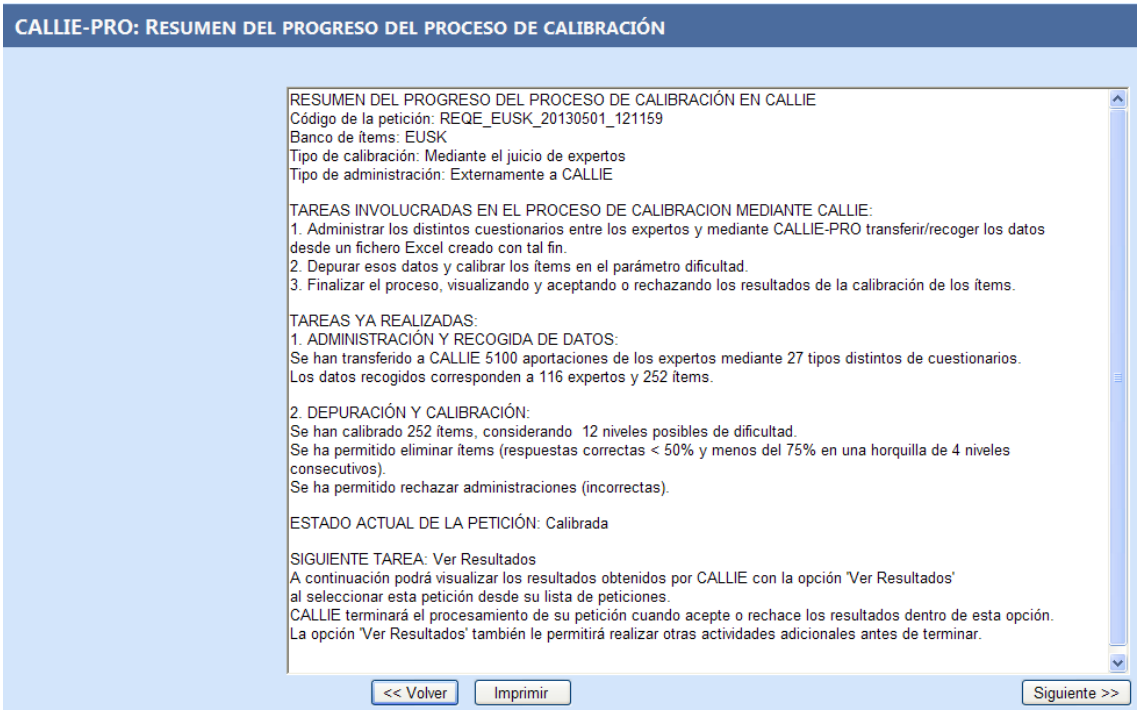


Figura 41 – CALLIE-PRO: Página ver resumen progreso.

Además, en toda petición no concluida y cuyo tipo de administración sea mediante CALLIE-MOODLE, se visualiza la opción *Ir al Moodle de CALLIE*. Su función es redirigir la aplicación a la Web de CALLIE-MOODLE (Figura 42) desde la

que el responsable puede acceder al curso asociado a su petición. El propio navegador utilizado permite volver a CALLIE-PRO.

Figura 42 – CALLIE-PRO: Web de CALLIE-MOODLE.

Además de las tres opciones mencionadas, existen otras opciones específicas para las peticiones no concluidas, que aparecen en la parte derecha de la Figura 43, y que dependen del estado en el que se encuentre la petición seleccionada. Estas opciones son las que se detallan en el plano de la Figura 38 e invocan las tareas *Recoger (más) datos* y *Depurar y Calibrar* de PRO-C. En concreto, las dos primeras filas que se muestran en la Figura 43 se corresponden con las opciones activas en dos peticiones en estado aceptada, la primera para una petición que no utiliza CALLIE-MOODLE como sistema de administración y la segunda para otra que sí lo utiliza. Las opciones de la tercera fila se corresponden con las de una petición en estado administrada. Una vez que la muestra ha sido depurada y calibrada por primera vez, la petición pasa a estado calibrada y se tiene acceso a la opción *Ver resultados*, que da acceso al resto de tareas de PRO-C.

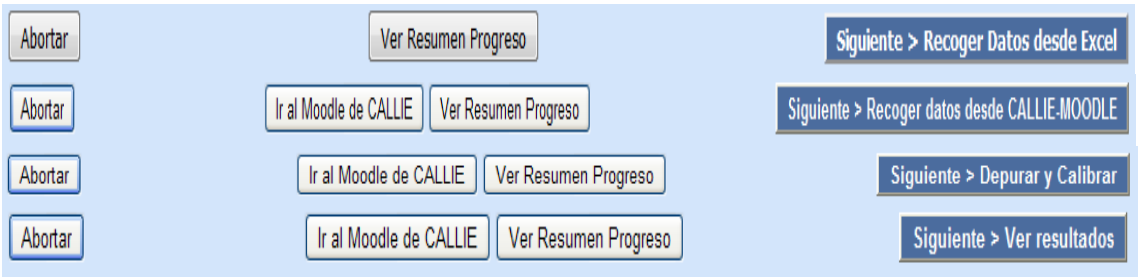


Figura 43 – CALLIE-PRO: Distintas opciones posibles para una petición no concluida.

La selección de la opción *Recoger Datos desde Excel*, lleva a una nueva página web (Figura 44) que ayuda al usuario a crear adecuadamente un fichero Excel con los datos conformes al MMAP y a transferirlo a CALLIE-EXPERT mediante una serie de pasos.

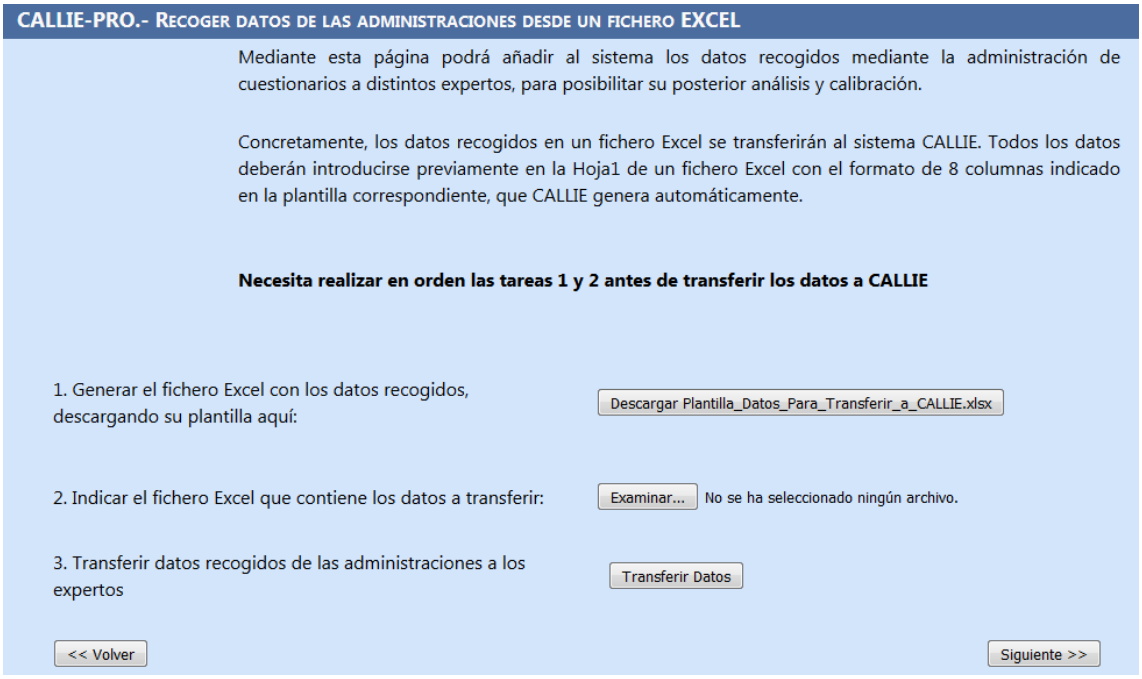


Figura 44 – CALLIE-PRO: Página generar Excel.

La selección de cualquiera de las dos opciones *Recoger datos desde CALLIE-MOODLE* o *Depurar y Calibrar* realiza la tarea correspondiente y cambia el estado de la petición, reflejándolo en la lista de peticiones.

La selección de la opción *Ver resultados* lleva a la segunda página principal de CALLIE-PRO – "Ver resultado de la calibración" – que ofrece un *resumen con los resultados finales* según el modelo de datos conforme al MMRE divididos en tres partes como se aprecia en la Figura 45.

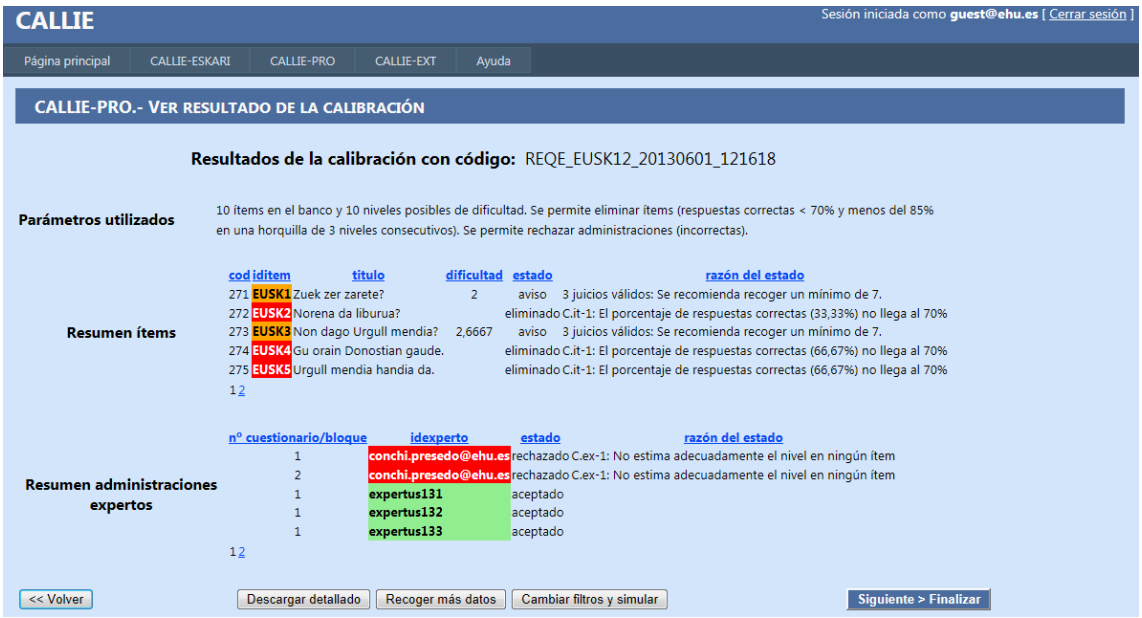


Figura 45 – CALLIE-PRO: Página ver resultado de la calibración.

En la zona superior del informe se indica el código de la petición y los parámetros especificados en el MMANCA y utilizados en los análisis. En la zona central superior se muestran los resultados finales de la calibración de cada ítem del banco. En la zona central inferior se visualiza el informe referido a las administraciones de los expertos que han participado. En ambos casos, las listas se pueden ordenar – de forma ascendente o descendente – por cualquier columna. Para facilitar la revisión de estos resultados, además de avisos textuales, se ha mantenido el código de colores para indicar aceptación (verde), aceptación aunque no cumple las recomendaciones (naranja) y rechazo o eliminación (rojo). En la parte inferior de la página, el botón << *Volver* regresa a la lista de peticiones y CALLIE-PRO ofrece otra serie de opciones que se describen a continuación.

La opción *Descargar detallado* consta de una serie de informes que contienen todos los datos y cálculos intermedios del modelo de resultados conforme al MMRE. Estos informes constituyen el **informe de resultados** de la herramienta y se generan automáticamente en distintas hojas de un libro MS Excel como el de la Figura 46.

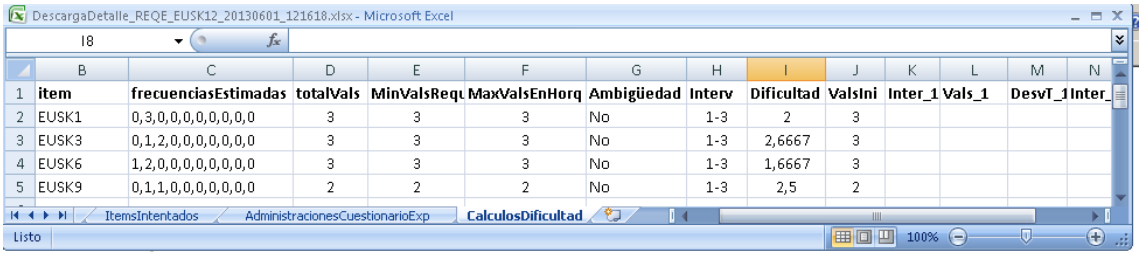


Figura 46 – CALLIE-PRO: Informe de resultados generado en la descarga MS Excel.

La primera hoja del informe de resultados – *DatosRecogidos* – contiene los datos recogidos de cada aportación de la muestra de partida con indicación de si la entrada es válida o ha sido rechazada junto con el motivo del rechazo. La segunda hoja – *AdministracionesCuestionarioExp* – guarda los resultados de los cálculos realizados para descartar expertos. Cada fila de esta hoja corresponde a una administración de un cuestionario a un experto. La tercera hoja – *ItemsIntentados* – almacena el detalle de los resultados obtenidos por CALLIE-EXPERT para los ítems del banco original contestados por los expertos. La cuarta hoja – *CalculosDificultad* – recoge las frecuencias recopiladas de los expertos para cada nivel a estimar, junto con las valoraciones utilizadas para estimar la dificultad y los cálculos relevantes en los casos de ítems con intervalos ambiguos. Como ejemplo, las tablas del anexo A4 muestran el contenido de estas tres últimas hojas para el caso Hezinet que se discutirán en la sección XI 2.1.

La opción *Recoger más datos* realiza la tarea Recoger (más) datos de PRO-C, desde el sistema de administración elegido.

La opción *Cambiar filtros y simular* realiza la tarea homónima de PRO-C, para lo cual dispone de una página (Figura 47) que permite cambiar el diseño de los análisis a aplicar sobre la muestra recogida, obteniendo los resultados en un fichero Excel.

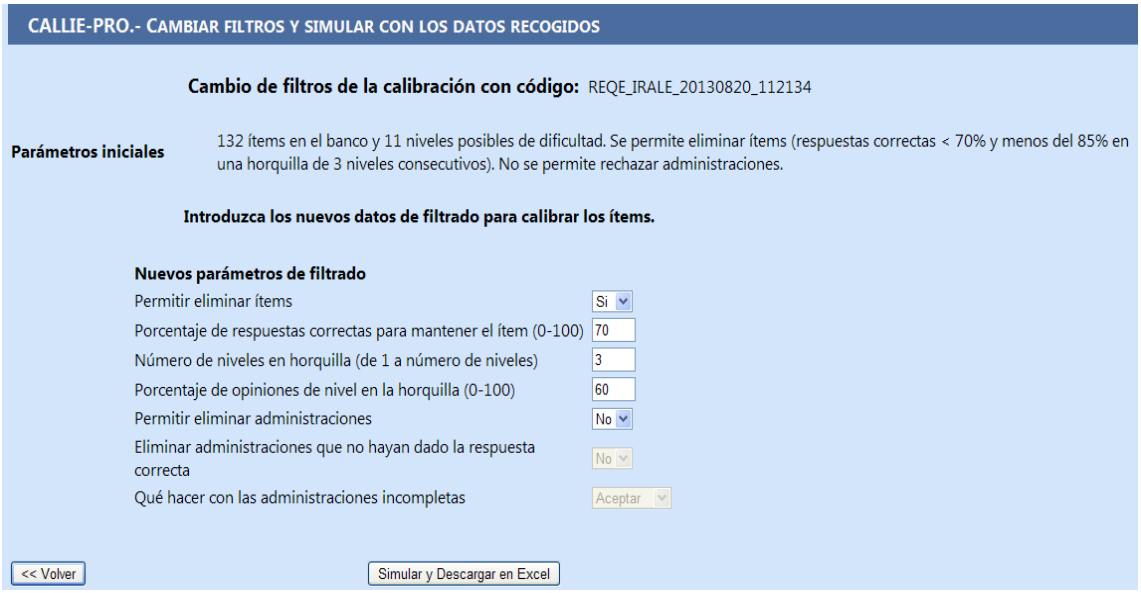


Figura 47 – CALLIE-PRO: Página cambio de filtros y simulación.

Por su parte, la opción *Finalizar* permite concluir la calibración invocando la tarea Aceptar/Rechazar resultados de PRO-C, que pasa el estado de la petición a Finalizada (directamente si es el tema EUSK de prueba). La Figura 48 muestra la pantalla que permite al responsable/creador del tema calibrado guardar o no los resultados de su calibración en el sistema CALLIE, ofreciéndole distintas opciones. Las dos primeras rechazan el almacenamiento de los resultados, aunque la segunda permite conservar el tema con sus ítems.

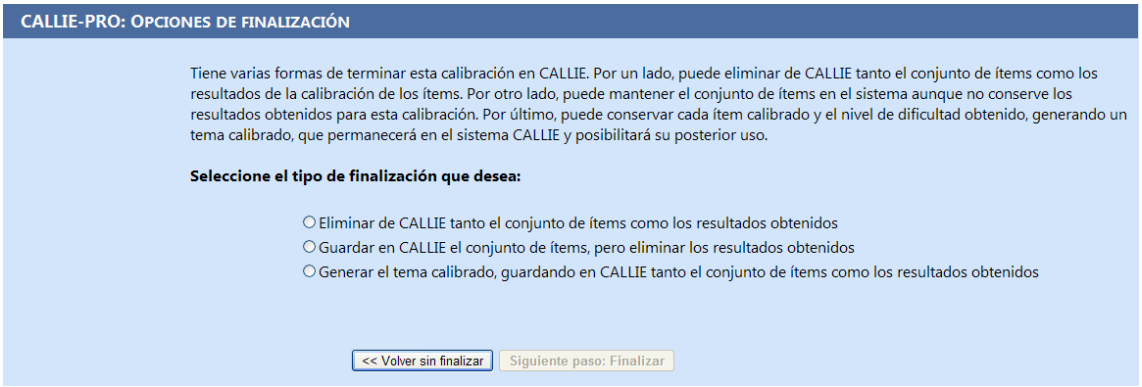


Figura 48 – CALLIE-PRO: Página opciones de finalización.

La tercera opción *Generar el tema calibrado* también pasa el estado de la petición a finalizada, pero almacena los ítems y sus valores de dificultad en el sistema CALLIE obteniéndose de este modo el banco de ítems calibrado. Como ejemplo, en la Figura 49 se muestra la pantalla que permite modificar dificultades manualmente, e incluso borrar ítems, antes de guardar los valores definitivos del tema calibrado en CALLIE.

Figura 49 – CALLIE-PRO: Página generar banco calibrado.

XI Evaluación de CALLIE-EXPERT

Este capítulo evalúa que el sistema creado es adecuado para su propósito, a través de una serie de pruebas con usuarios para comprobar que cada uno de sus componentes cumple su función sin problemas y el conjunto también (como consecuencia de lo anterior). En estas pruebas se ha empleado CALLIE-EXPERT para reproducir la calibración del banco de ítems de Hezinet y para realizar una calibración con idénticas características que la calibración manual previa del banco de ítems preparado por el servicio de formación de euskera para el profesorado IRALE (IRakasleen ALfabetatze eta Euskalduntzea) y después replicarla.

El banco de ítems de Hezinet se utiliza para posicionar estudiantes de euskera en su nivel apropiado dentro de dicho sistema. Consta de 252 ítems de selección múltiple para la evaluación del conocimiento gramatical de la lengua vasca. El banco no tiene categorías ni niveles previos. La estructura de cada ítem es básica: el enunciado de la pregunta y cuatro posibles respuestas, todas textuales, de las cuales solamente una es correcta. El objetivo de esta calibración era obtener la dificultad de cada ítem en términos de los 12 niveles definidos por HABE (1984). Este banco ha sido calibrado vía expertos manualmente por Arruabarrena (2008). También ha sido calibrado mediante TRI manualmente por López-Cuadrado (2008) y automatizadamente por Armendariz (2014).

El banco de ítems preparado por IRALE se utiliza para conocer el nivel de comprensión lectora en lengua vasca de estudiantes de euskera cuyas edades van desde los 11 hasta los 18 años. Consta para tal fin de 132 ítems de selección múltiple, 7 categorías posibles y 3 niveles previos. La estructura de cada ítem es básica: un texto a leer, una pregunta y cuatro posibles respuestas, de las cuales solamente una es correcta. El objetivo de esta calibración era obtener la dificultad de cada ítem en términos de los 11 niveles definidos por el Marco Común Europeo de Referencia para las lenguas o CEFR (Council_of_Europe, 2001) para la lectura de un idioma. La calibración vía expertos de este banco constituyó el primer paso para obtener, finalmente, un banco calibrado de ítems que ha sido incluido en un TAI operativo llamado HIZEBA (http://irakurketa.hizeba.eu/) que sigue en uso en la actualidad.

Para evaluar cada componente del sistema CALLIE-EXPERT con usuarios, se han realizado dos experimentos sucesivos en el tiempo: el experimento 1 y el experimento 2. Mediante el experimento 1, descrito en el anexo A2, se ha llevado a cabo la calibración automatizada vía expertos del banco de ítems IRALE, que ya había sido calibrado manualmente mediante colaboración de miembros del departamento de Lenguajes y Sistemas Informáticos de la UPV/EHU con los responsables de IRALE. También se ha llevado a cabo una evaluación preliminar de los principales componentes de CALLIE-EXPERT. Mediante el experimento 2, descrito en el anexo A3, se completa la prueba de componentes de CALLIE-EXPERT. Además, su ejecución ha permitido replicar las calibraciones de los ítems de Hezinet y de IRALE puesto que en el experimento se utilizaron datos obtenidos en las calibraciones originales.

Como consecuencia del experimento 2, se han reproducido mediante CALLIE-EXPERT las dos calibraciones manuales ya mencionadas: la realizada por Arruabarrena (2008) y la realizada sobre el banco de ítems IRALE, para comprobar que los resultados que proporciona la herramienta son acordes a los obtenidos previamente en estas dos calibraciones.

Además, se han llevado a cabo una serie de pruebas globales sobre la aplicación Web CALLIE con otros usuarios, en el contexto de la evaluación del módulo EXT.

Por último, se han desarrollado una serie de pruebas de integración con CALLIE-TRI.

En las siguientes secciones se detalla cada una de estas pruebas y se discuten los resultados obtenidos.

XI 1 Pruebas de componentes con usuarios

Para probar cada componente de CALLIE-EXPERT se ha contado con dos grupos de usuarios. En el experimento 1 ha participado un grupo de 8 profesores de la UPV/EHU (Exp1, Exp2, Exp3, Exp4, Exp5, Exp6, Exp7 y Exp8) entre los cuales los dos primeros actúan además como responsables de la calibración automatizada del banco de ítems de IRALE (Respons1 y Respons2) y todos ellos como expertos en la misma calibración. En el experimento 2 ha participado un grupo de 5 alumnos universitarios del grado de ingeniería informática (Alu1, Alu2, Alu3, Alu4 y Alu5). El primer grupo son usuarios potenciales de CALLIE, conocedores de la plataforma Moodle en el rol de profesor (pero no necesariamente a nivel de alumno en cuanto a cuestionarios y su envío), de temas de evaluación y calibración con poca experiencia en herramientas informáticas, y el segundo son sujetos con pocas posibilidades de utilizar CALLIE, sin conocimientos de temas de evaluación y calibración, pero con experiencia en distintas herramientas informáticas entre ellas la plataforma Moodle en el rol de alumno.

Los dos experimentos han estado formados por cuatro pruebas sucesivas. En cada experimento, cada una de estas 4 pruebas ha servido para evaluar por orden las siguientes funcionalidades y componentes de CALLIE-EXPERT: creación del banco e introducción de ítems con ESKARI (prueba 1); diseño/especificación del experimento con ESKARI (prueba 2); administración de ítems a los expertos con CALLIE-MOODLE y/o importación de datos de administraciones de ítems externas para PRO-C preparando un libro Excel (prueba 3); y recogida de datos, realización del análisis y calibración y discusión de los resultados con PRO-C (prueba 4).

Los siguientes epígrafes resumen en qué ha consistido cada una de estas cuatro pruebas de un modo global, quiénes han sido sus participantes, las funcionalidades que han sido abarcadas en total a través de las tareas involucradas, y detallan los resultados obtenidos tras los dos experimentos. Tras ello, en un último apartado se unifican los resultados anteriores para indicar los resultados finales globales para cada componente de CALLIE-EXPERT junto a sus puntuaciones.

XI 1.1 Pruebas de introducción de ítems con ESKARI

Para probar la creación del banco e introducción de ítems con ESKARI se ha pedido a 7 personas, con distintos grados de conocimientos en informática y en temas de evaluación y calibración, que generen ítems en CALLIE utilizando ESKARI. Concretamente, se ha pedido a los dos profesores de la UPV/EHU que actuaban como responsables que creen un nuevo banco de ítems y que introduzcan los 132 ítems de su banco clasificados en categorías y niveles previos y al grupo de alumnos se les ha pedido que introduzcan 5 ítems de selección múltiple con la utilidad que automatiza el formato IMS QTI y otros 5 importando el ítem en formato IMS QTI. La Tabla 10 resume las tareas involucradas en estas pruebas y los resultados relevantes.

Tareas Introducción ítems	Respons1	Respons2	Alu1	Alu2	Alu3	Alu4	Alu5	Realización / punt. media
Creación banco	IRALE	IRALE	-	-	-	-	-	**100%**
Selección banco existente	Ok	Ok	Ok	Mal	Ok	Ok	Ok	**85,71%**
Introducción ítems	Ok	Ok	Ok	Ok	Ok	Ok	Mal	**85,71%**
Items introducidos mediante utilidad	66	66	5	5	5	5	5	**100%**
Items introducidos mediante importación	-	-	5	5	5	5	0	**80%**
Puntuación Introducción ítems	**8**	**8**	**7**	**6**	**6**	**7**	**6**	**6,88**

Tabla 10 – Resumen de las pruebas realizadas para la introducción de ítems.

Como se refleja en la tabla anterior, estas pruebas de introducción de ítems han abarcado la utilización del paso 1 del menú de CALLIE-ESKARI para: (1) crear un tema con categorías y niveles previos, (2) seleccionar un tema ya existente (3) introducir muchos y pocos ítems tanto en un tema con categorías y niveles previos como en otro sin ellos, (4) aplicar los distintos métodos de introducción de ítems disponibles en CALLIE: la utilidad presente a través de la opción *Generar QTI* de la interfaz o la importación de ítems en IMS QTI que obliga a tenerlos previamente en este formato (Figura 25 de la memoria).

Todos los participantes han considerado que el componente es adecuado para su propósito y funciona correctamente, les gusta que se pueda visualizar el banco a medida que se introduce cada ítem, pero algunos de ellos echan en falta la posibilidad de introducir varios ítems a la vez. En cuanto a los resultados, siempre que el participante ha podido elegir ha introducido los ítems con la utilidad. Solamente hubo un participante que no finalizó su prueba porque falló al crear los ítems en formato IMS QTI, puesto que no entendió cómo debía hacerlo a pesar de consultar la ayuda disponible. Todos los demás (85,71%) hicieron correctamente su tarea dentro del tiempo asignado a la misma. Los participantes otorgaron a este componente una puntuación media de 6,88.

XI 1.2 Pruebas de diseño del experimento con ESKARI

Para probar el diseño del experimento con ESKARI se ha pedido a 7 personas, con distintos grados de conocimientos en informática y en temas de evaluación y

calibración, que reproduzcan el diseño de los dos experimentos de calibración que se han considerado hasta el momento: Hezinet e IRALE. Concretamente, se ha pedido a los dos profesores de la UPV/EHU que actuaban como responsables que reproduzcan el diseño para la calibración de IRALE y al grupo de alumnos se les ha pedido que reproduzcan los dos. Para ello se les ha proporcionado el banco de ítems correspondiente ya creado y las características necesarias para poder diseñar y administrar los cuestionarios con CALLIE-EXPERT y se les ha pedido que creen el experimento para calibrar sus ítems vía expertos utilizando las mismas opciones empleadas en el diseño original de esas dos calibraciones. Las opciones que debían elegir en cada caso aparecen detalladas en la Tabla 11.

PASO#	CARACTERÍSTICA DE DISEÑO PEDIDA	VALOR PARA IRALE	VALOR PARA HEZINET
Paso 1	Conjunto (banco) de ítems a calibrar	IRALE	EUSK (seleccionando ítems de prueba o el tema EUSK)
Paso 2	Tipo de calibración a llevar a cabo	Expertos	Expertos
Paso 3	Tipo de reparto inicial de los ítems	Continuo	Alterno (por defecto)
	Administración de ítems mediante CALLIE-MOODLE	Sí	No (por defecto)
	Datos de acceso experto	expXXX	-
	Tamaño homogéneo de ítems en cada cuestionario	No	Sí (por defecto)
	Nº Cuestionarios, bloques y asignación de ítems	Según documentación proporcionada (ver Anexo A2)	6 cuestionarios, 42 ítems por cuestionario y asignación alterna (por defecto)
	Número de expertos disponibles	8	116
Paso 4	Num. niveles de dificultad (mínimo 2)	11	12
	Permitir dejar el nivel en blanco	Sí	No (por defecto)
	Obligar a dar la respuesta correcta	No	Sí (por defecto)
	% respuestas correctas para mantener el ítem (0-100)	70 (por defecto)	50
	Núm. de niveles en horquilla (de 1 a número de niveles)	3 (por defecto)	4
	% de opiniones de nivel en la horquilla (0-100)	85 (por defecto)	75
	Permitir eliminar administraciones	No	Sí (por defecto)

Tabla 11 – Características de cada una de las calibraciones a diseñar.

Por su parte, la Tabla 12 resume las tareas involucradas en esta prueba y los resultados relevantes.

Tareas diseño	Respons1	Respons2	Alu1	Alu2	Alu3	Alu4	Alu5	Realización / punt. media
Pasos 1 a 4 Hezinet	-	-	Sel. banco	Sel. banco	Ítems prueba	Sel. banco	Ítems prueba	**100%**
Pasos 1, 2 y 4 IRALE	Ok	Ok	Ok	Ok	Ok	Ok	Ok	**100%**
Reparto inicial IRALE	Continuo	Continuo	Continuo	Alterno	Continuo	Alterno	Alterno	**57,14%**
Resto paso 3 IRALE	Ok	Ok	Ok	Ok	Ok	Ok	Mal	**85,71%**
Confirmar Petición	Ok	Ok	Ok	Ok	Ok	Ok	Ok	**100%**
Puntuación Menú	**9**	**9**	**8**	**8**	**9**	**9**	**9**	**8,71**

Tabla 12 – Resumen de resultados de las pruebas para el diseño del experimento.

Como se refleja en las dos tablas anteriores, estas pruebas de diseño del experimento han abarcado la utilización del menú de cuatro pasos de CALLIE-ESKARI

para: (1) seleccionar un banco de ítems ya creado, (2) calibrar un banco de ítems mediante el juicio de expertos con varios de los supuestos que permite la herramienta como son, por ejemplo, emplear expertos disponibles escasos o de sobra, utilizar una administración de ítems con CALLIE-MOODLE o una administración de otro tipo, usar un diseño de cuestionarios homogéneos o usar un diseño no homogéneo, (3) elegir el tipo de reparto inicial más adecuado para asignar los ítems a los cuestionarios, (4) tomar distintas decisiones de filtrado y administración y (5) ver resumen, corregir y confirmar petición comprobando que está en CALLIE-PRO.

Todos los participantes han considerado que el componente es adecuado para su propósito y funciona correctamente. Han valorado positivamente el cálculo automático de datos y las alertas sobre posibles errores que aparecieron durante el diseño de la calibración de los ítems de IRALE. No hubo comentarios negativos. En cuanto a los resultados, todo participante excepto un alumno realizó su tarea correctamente (85,71%) y dentro del tiempo establecido para la prueba. El único alumno que no consiguió terminar la tarea, fracasó debido a una mala distribución de los ítems en los distintos bloques para cuestionarios no homogéneos, puesto que no los redistribuyó acorde a la tabla proporcionada. Los participantes otorgaron a este componente una puntuación media de 8,71.

XI 1.3 Pruebas con CALLIE-MOODLE

Para probar la administración de ítems con CALLIE-MOODLE se ha pedido a los 8 integrantes del grupo de profesores de la UPV/EHU para la calibración automatizada del tema IRALE (inexpertos en la funcionalidad de cuestionarios de la plataforma Moodle) que realicen la administración de los ítems de la calibración IRALE con este componente. Con objeto de evaluar la funcionalidad de la plataforma CALLIE-MOODLE tanto desde el punto de vista del responsable como desde el punto de vista del experto, la prueba se ha dividido en dos partes. Por un lado, dos personas (los responsables) han conducido la administración de los ítems y han dado un plazo para dar por concluida esa administración. Por otro lado, los 8 integrantes del grupo han participado como expertos cumplimentando y enviando los cuestionarios. Para ello, se les ha proporcionado acceso a un curso en CALLIE-MOODLE asociado a la calibración diseñada por IRALE tal y como lo generaría la propia herramienta CALLIE-EXPERT.

La Tabla 13 resume las tareas involucradas para los responsables en esta prueba y los resultados relevantes.

Tareas administración responsables	Respons1	Respons2	Realización / punt. media
Tareas básicas (avisos cumplimentación, administración, revisión datos)	Moodle	Moodle	**100%**
Tareas no básicas (instrucciones, resolución de consultas)	email, teléfono	email, teléfono	**100%**
Puntuación CALLIE-MOODLE Responsables	**9**	**9**	**9**

Tabla 13 – Resumen de la prueba CALLIE-MOODLE parte responsables de la calibración.

Por su parte, la Tabla 14 resume las tareas involucradas para los expertos en esta prueba y los resultados relevantes.

Tareas administración expertos	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7	Exp8	Realización / punt. media
Rellenar cuestionario	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	**100%**
Enviar cuestionario	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	**100%**
Puntuación CALLIE-MOODLE Expertos	**9**	**9**	**8**	**8**	**8**	**8**	**9**	**7**	**8,25**

Tabla 14 – Resumen de la prueba CALLIE-MOODLE parte expertos participantes.

Como se refleja en las dos tablas anteriores, estas pruebas de administración de ítems han abarcado la utilización de CALLIE-MOODLE para: (1) realizar las tareas básicas del responsable en la administración de ítems como avisos sobre la cumplimentación de cuestionarios, la propia administración de los ítems y la revisión de los datos aportados por los expertos, (2) ofrecer al responsable la posibilidad de usar Moodle en otras tareas no básicas como adjuntar documentación con instrucciones, enviar y resolver consultas, etc., y (3) realizar las tareas básicas del experto en la administración de ítems: rellenar y enviar los cuestionarios dentro del plazo definido.

Todos los participantes han considerado que CALLIE-MOODLE es una plataforma adecuada tanto para administrar los ítems y seguir el progreso de los expertos como para la cumplimentación de los cuestionarios. Se registraron algunos problemas, que no eran debidos a un mal funcionamiento del componente sino al desconocimiento de las distintas opciones de la plataforma. A pesar de ello, todo participante realizó sus tareas correctamente y dentro del tiempo establecido para la prueba. Sin embargo, la revisión de la utilización o no de Moodle en las tareas que realizaron los responsables reveló que solamente se empleó Moodle cuando era estrictamente necesario, seguramente debido a su inexperiencia con esta plataforma. Los participantes otorgaron a este componente una puntuación media de 9 para las tareas de administración de los responsables y una puntuación media de 8,25 para las de los expertos.

XI 1.4 Pruebas con Excel para PRO-C

Para probar la recogida de datos para PRO-C desde un sistema de administración de ítems distinto a CALLIE-MOODLE se ha pedido a 5 personas inexpertas en temas de calibración (los cinco alumnos) que transformen los datos de una muestra de datos procedente de una administración externa para recogerla mediante CALLIE-PRO. Para ello se les ha proporcionado un documento con los datos recabados de una administración con 11 niveles que siguen la escala dada por el CEFR y cuya petición de calibración está preparada en CALLIE-EXPERT para que pueda recogerlos. La Tabla 15 resume las tareas involucradas en esta prueba y los resultados relevantes.

Tareas recogida datos Excel	Alu1	Alu2	Alu3	Alu4	Alu5	Realización / punt. media
Encontrar la ayuda	Ok	Ok	Ok	Ok	Mal	**80%**
Modificar escalas	Ok	Mal	Ok	Ok	Mal	**60%**
Formato	Ok	Ok	Ok	Ok	Mal	**80%**
Nombre y ubicación	Ok	Ok	Ok	Ok	Ok	**100%**
Recoger datos desde Excel a la petición de prueba	Ok	Mal	Ok	Ok	Mal	**60%**
Puntuación PRO-C Excel	**7**	**6**	**6**	**7**	**6**	**6,4**

Tabla 15 – Resumen de la prueba con Excel.

Como se refleja en la tabla anterior, estas pruebas de recogida de datos con Excel han abarcado la utilización de las opciones de la interfaz inicial de PRO-C para: (1) encontrar la ayuda sobre el formato que debe tener el fichero Excel a recoger, el nombre y la ubicación en la que se debe colocar, (2) cambiar la escala de niveles de dificultad a la requerida por CALLIE-EXPERT y (3) recoger los datos formateados.

Todos los participantes han considerado que el componente es adecuado para su propósito y funciona correctamente. Han valorado positivamente que exista una utilidad de estas características pero la juzgan excesivamente rudimentaria a la hora de ayudar a crear y cargar el fichero Excel en el sistema. De hecho, solamente tres de los cinco participantes (60%) terminaron estas tareas correctamente y dentro del tiempo establecido para la prueba. Los participantes otorgaron a este componente una puntuación media de 6,4.

Debido a esta media tan baja y a los malos resultados obtenidos con los usuarios, una vez concluidas estas pruebas, en las que se usaba la interfaz inicial, se mejoró este componente como se ha indicado en la sección IX 3.5.

XI 1.5 Pruebas con PRO-C

Para probar el funcionamiento del componente PRO-C del sistema CALLIE-EXPERT mediante la interfaz CALLIE-PRO se ha pedido a 7 personas, con distintos grados de conocimientos en informática y en temas de evaluación y calibración, que calibren y discutan los resultados. Concretamente, se ha pedido a los dos profesores de la UPV/EHU que actuaban como responsables y a los cinco alumnos que tomen los datos de CALLIE-MOODLE, que calibren y discutan los resultados partiendo del diseño/especificación de dos peticiones de calibración en estado aceptada ya generadas con ESKARI y de los datos dados por los expertos en sus administraciones. Para ello, se les ha proporcionado la petición inicial correspondiente a IRALE introducida con CALLIE-ESKARI en estado aceptada y lista para recoger los datos de sus administraciones desde un curso de CALLIE-MOODLE. A continuación, se les ha pedido a estas 7 personas que discutan los resultados iniciales obtenidos para IRALE y que obtengan resultados alternativos sobre los mismos datos. Por otro lado, a los 5 alumnos se les ha proporcionado la petición inicial correspondiente a Hezinet introducida con CALLIE-ESKARI en estado aceptada y se les ha pedido que suban los datos desde un libro Excel que contiene las administraciones Hezinet y realicen los pasos de la calibración hasta obtener el informe de resultados.

La Tabla 16 resume las tareas involucradas en esta prueba sobre el progreso del proceso y los resultados relevantes.

Tareas	Respons1	Respons2	Alu1	Alu2	Alu3	Alu4	Alu5	Realización / punt. media
Recoger datos desde Excel	-	-	Ok	Ok	Ok	Ok	Ok	100%
Recoger datos desde Moodle	Ok	Ok	Ok	Ok	Ok	Ok	Ok	100%
Depurar y calibrar	Ok	Ok	Ok	Ok	Ok	Ok	Ok	100%
Ver resultados	Ok	Ok	Ok	Ok	Ok	Ok	Ok	100%
Puntuación PRO-C progreso calibración	9	9	9	9	9	10	8	9

Tabla 16 – Resumen de la prueba PRO-C sobre el progreso del proceso.

Por su parte, la Tabla 17 resume las tareas involucradas en esta prueba sobre la discusión de resultados y los resultados relevantes.

Tareas	Respons1	Respons2	Alu1	Alu2	Alu3	Alu4	Alu5	Realización / punt. media
Ver resultados detallados	**Ok**	**Ok**	**Ok**	**Ok**	**Ok**	**Ok**	**Ok**	**100%**
Simular	**Ok**	**Ok**	**Ok**	**Ok**	**Ok**	**Ok**	**Ok**	**100%**
Responder preguntas	**-**	**-**	**Ok**	**Ok**	**Ok**	**Ok**	**Ok**	**100%**
Puntuación PRO-C discusión resultados	**9**	**9**	**8**	**9**	**8**	**8**	**8**	**8,43**

Tabla 17 – Revisión de la prueba PRO-C sobre la discusión de resultados.

Como se refleja en la Tabla 16 y la Tabla 17, las pruebas del componente PRO-C han abarcado la utilización de las distintas opciones de su interfaz CALLIE-PRO para: (1) recoger los datos de las administraciones desde el fichero Excel proporcionado y desde el curso de CALLIE-MOODLE asociado a la petición, (2) analizar y calibrar, (3) ver resultados detallados, (4) discutir resultados y (5) simular resultados con otros filtros. Cabe destacar que las dos primeras tareas son totalmente automáticas para una petición aceptada.

Todos los participantes han considerado que PRO-C es un componente adecuado y que funciona correctamente, tanto para controlar el progreso de la calibración como para discutir los resultados que se obtienen. Han valorado positivamente la información que proporciona la herramienta en cada momento, la automatización de los sucesivos pasos una vez aceptada la petición de calibración y la posibilidad de ordenar los resultados. El único inconveniente que se ha registrado es que la lista de peticiones no se puede ordenar, lo que no es un problema grave porque un usuario normalmente no tendrá tantas calibraciones en curso como para tener que ordenarlas. Los participantes otorgaron a este componente una puntuación media de 9 para el control del progreso de la calibración y una puntuación media de 8,43 para la discusión de los resultados.

XI 1.6 Resultados globales para los componentes

En cuanto a las opiniones de los participantes sobre las funcionalidades de los distintos componentes de CALLIE-EXPERT para la calibración de ítems vía expertos, los dos grupos consideraron que todos son adecuados para este propósito y que funcionan correctamente.

Los resultados globales que se obtuvieron para las funcionalidades de CALLIE-EXPERT evaluadas se reflejan en la Figura 50. Se obtuvo una media de 6,88 como puntuación para la introducción de ítems mediante ESKARI (*ESKARI Introducción ítems*), 8,71 para el diseño del experimento mediante el menú de ESKARI (*ESKARI Menú*), 9 para las tareas de administración de los responsables mediante CALLIE-MOODLE (*CALLIE-MOODLE Responsables*), 8,25 para las tareas de los expertos durante la administración de ítems mediante CALLIE-MOODLE (*CALLIE-MOODLE Expertos*), 6,4 para la recogida de datos desde Excel mediante PRO-C (*PRO-C Excel*), 9 para el control y seguimiento del progreso del proceso de calibración mediante PRO-C (*PRO-C Progreso calibración*) y 8,43 para la discusión de los resultados obtenidos mediante PRO-C (*PRO-C Discusión resultados*).

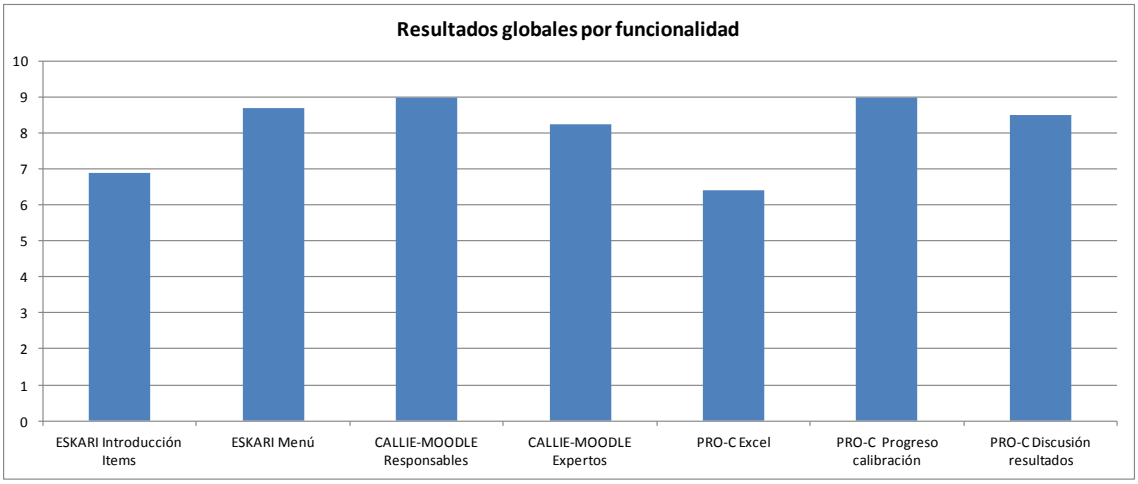


Figura 50 – Resultados globales por funcionalidad.

Los resultados por funcionalidad y grupo de participantes para los distintos componentes de CALLIE-EXPERT se recogen en la Tabla 18. Aunque no existen grandes diferencias entre ambos grupos, en todos los casos los usuarios conocedores de los temas de evaluación y calibración han dado una puntuación más alta.

	GrupoIRALE	GrupoAlumnos
ESKARI Introducción ítems	8	6,4
ESKARI Menú	9	8,6
CALLIE-MOODLE Responsables	9	-
CALLIE-MOODLE Expertos	8,5	-
PRO-C Excel	-	6,4
PRO-C Progreso calibración	9	9
PRO-C Discusión resultados	9	8,2

Tabla 18 – Puntuaciones medias por funcionalidad y grupo para CALLIE-EXPERT.

Como resultados globales por componente de CALLIE-EXPERT (Figura 51) se ha obtenido una puntuación media de 7,8 para ESKARI, de 8,62 para CALLIE-MOODLE y de 7,97 para PRO-C.

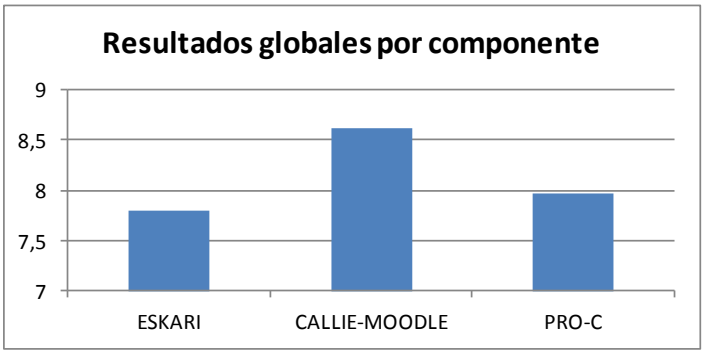


Figura 51 – Resultados globales por componente.

XI 2 Resultados obtenidos por CALLIE-EXPERT

Tanto la autora como los cinco alumnos participantes en el experimento 2, han replicado con CALLIE-EXPERT las calibraciones de los dos bancos de ítems citados a lo largo de la memoria, con lo que se han obtenido 6 réplicas de cada calibración, todas ellas con idénticos resultados. En este apartado se reflejan estos resultados obtenidos por la herramienta. Los informes de resultados detallados obtenidos por CALLIE-EXPERT en estas dos calibraciones se pueden consultar en los Anexos A4 y A5 respectivamente.

XI 2.1 Réplica de la calibración de los ítems de Hezinet

Como en este caso se trataba de reproducir una calibración ya realizada, se hizo la especificación de la calibración de los ítems aplicando las mismas restricciones que se impusieron en la calibración original. Por otro lado, los ítems del banco, los datos de los expertos y la muestra recogida con las valoraciones de estos expertos sobre los ítems se hallaban almacenados en una base de datos MS Access, por lo que dicho contenido se tomó para alimentar al sistema CALLIE.

Al preparar el experimento 2 se ha creado en CALLIE-EXPERT un tema denominado EUSK con los 252 ítems del banco Hezinet. Asímismo, mediante el componente ESKARI de la herramienta se ha diseñado el experimento en CALLIE-EXPERT y se ha generado una petición de calibración utilizando el menú de cuatro pasos de CALLIE-ESKARI con el código REQE_EUSK_20130501_121159 y las características originales de Hezinet que se muestran en la Tabla 19.

PASO#	CARACTERÍSTICA DE DISEÑO	VALOR PARA HEZINET
Paso 1	Conjunto de ítems a calibrar	EUSK
Paso 2	Tipo de calibración a llevar a cabo	Expertos
Paso 3	Tiempo máximo por cuestionario (minutos)	45
	Tiempo medio por ítem y parámetro (segundos)	60
	Tipo de reparto inicial de los ítems	Alterno
	Administración de ítems mediante CALLIE-MOODLE	No
	Datos de acceso experto	-
	Tamaño homogéneo de ítems en cada cuestionario	Sí
	Número de cuestionarios diferentes	6
	Número de ítems por cuestionario	42
	Número de bloques	-
	Número de ítems por bloque	-
	Número de expertos disponibles	116
	Porcentaje estimado de abandono expertos (0-100)	0
Paso 4	Método para estimar la dificultad	M.dif
	Num. niveles de dificultad (mínimo 2)	12
	Permitir dejar el nivel en blanco	No
	Pedir la respuesta correcta	Sí
	Obligar a dar la respuesta correcta	Sí
	Permitir aportaciones propias	Sí

PASO#	CARACTERÍSTICA DE DISEÑO	VALOR PARA HEZINET
	Permitir eliminar ítems	Sí
	% respuestas correctas para mantener el ítem (0-100)	50
	Núm. de niveles en horquilla (de 1 a número de niveles)	4
	% de opiniones de nivel en la horquilla (0-100)	75
	Permitir eliminar administraciones	Sí
	Eliminar adms. con respuesta incorrecta	Sí
	Qué hacer con las administraciones incompletas	Aceptar

Tabla 19 – Características de diseño para la réplica de Hezinet.

Durante la introducción de estos valores, CALLIE-EXPERT no generó ninguna alerta puesto que todos los valores cumplían las recomendaciones del sistema.

También se creó un libro Excel, siguiendo el formato adecuado para PRO-C, con los datos almacenados en el experimento original durante la fase de administración de los ítems a los expertos. Estos datos en Excel se transfirieron a CALLIE mediante la opción *Recoger datos desde Excel* de CALLIE-PRO. La muestra inicial para CALLIE-EXPERT quedó caracterizada por 5100 aportaciones, 252 ítems y 116 expertos.

Tras alimentar el sistema CALLIE con los datos proporcionados por los expertos, mediante la opción *Depurar y Calibrar* de CALLIE-PRO, se realizó el análisis de los datos de la muestra y la calibración de los ítems utilizando el workflow CA generado en la definición del experimento. Concretamente, para el análisis de datos se filtró la muestra aplicando en orden *C.ex-1* y *C.it-1* una sola vez, y se aplicaron reiteradamente los criterios *C.it-2* y *C.ex-2* en ese orden hasta estabilizar los resultados. A continuación, se calibró la dificultad de los ítems no descartados aplicando M.dif. Después, con la opción *Descargar detallado*, CALLIE-EXPERT generó el fichero Excel con las cuatro hojas del informe de resultados, cuyos datos más relevantes se han transcrito en las distintas tablas del anexo A4.

La Tabla 20 ilustra los criterios empleados y la evolución que sufrió la muestra de datos a lo largo del análisis de datos. En esta tabla, *m* es el número de entradas/aportaciones válidas de la muestra, *n* el número de ítems y *e* el número de expertos que se mantuvieron después de aplicar cada uno de los filtros considerados.

Filtro	Descripción	Aportaciones eliminadas	Ítems eliminados	Expertos eliminados	Muestra resultante
#1	Recogida datos	0	0	0	m=5100, n=252, e=116
#2	C.ex-1	577	0	1	m=4523, n=252, e=115
#3	C.it-1	115	7	0	m=4408, n=245, e=115
#4	C.it-2 (1ª iteración)	972	56	0	m=3436, n=189, e=115
#5	C.ex-2 (1ª iteración)	126	0	4	m=3310, n=189, e=111
#6	C.it-2 (2ª iteración)	75	5	0	m=3235, n=184, e=111
#7	C.ex-2 (2ª iteración)	0	0	0	m=3235, n=184, e=111

Tabla 20 – Evolución de la muestra durante el análisis de CALLIE-EXPERT para Hezinet.

La aplicación del criterio *C.ex*-1 mermó el tamaño de la muestra en 577 aportaciones (169 fuera de rango y 408 con varios niveles). Además descartó un experto *sin estimaciones válidas*, el experto 234, dando lugar a una nueva muestra con 4523

aportaciones, 252 ítems y 115 expertos. Seguidamente, se aplicó el criterio *C.it-1* y se identificaron 7 ítems que no superaban el umbral fijado del 50%, en concreto los ítems con códigos 16, 25, 151, 205, 234, 239 y 249. Como consecuencia, el tamaño de la muestra anterior se redujo en 115 aportaciones hasta 4408 entradas, siendo las características de la muestra resultante 4408 aportaciones, 245 ítems y 115 expertos. A esta última muestra se le aplicaron de *manera iterativa* los otros dos filtros en el orden C.it-2 y C.ex-2.

En una primera iteración, este proceso repetitivo *comenzó aplicando C.it-2*, lo que significó retirar del banco 56 ítems (un 23% de los ítems del banco) y sus respectivas entradas en la muestra (en total 972), mermando la muestra a 3436 aportaciones, 189 ítems y 115 expertos. En este punto se produjo la única diferencia con el experimento original en el que se eliminaron 6 ítems menos (Tabla 21), lo que fue debido al cálculo del mínimo de frecuencias necesario en la horquilla. Mientras que Arruabarrena (2010) redondeó el resultado al entero más próximo, CALLIE-EXPERT redondea siempre al siguiente entero por exceso.

ITEM	RAZÓN C.it-2 (6 ítems)
10	C.it-2: Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
56	C.it-2:Máximo valoraciones en horquilla (10) menor que el mínimo requerido (11)
115	C.it-2:Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
209	C.it-2:Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
235	C.it-2:Máximo valoraciones en horquilla (12) menor que el mínimo requerido (13)
237	C.it-2:Máximo valoracionDespuéses en horquilla (13) menor que el mínimo requerido (14)

Tabla 21 – Ítems extra retirados por el filtro C.it-2 en CALLIE-EXPERT.

El siguiente paso de esta primera iteración consistió en aplicar el criterio *C.ex-2*, que consideró inaceptables las aportaciones de los expertos 26, 91, 230 y 231, al obtener éstos una tasa de aciertos inferior al 75%. La retirada de dichas entradas (un total de 126) dio lugar a una muestra depurada con 3310 aportaciones, 189 ítems y 111 expertos.

En una segunda iteración, el criterio *C.it-2* descartó los ítems 11, 43, 153, 190 y 194, y dejó la muestra con 3235 aportaciones, 184 ítems y 111 expertos. A continuación, la nueva aplicación del criterio *C.ex-2* no modificó la muestra anterior, dando así por finalizada la depuración de la muestra tras dos iteraciones.

Finalizado el análisis de los datos, el workflow CA llevó a cabo la calibración de los ítems no descartados y obtuvo los 184 niveles de dificultad correspondientes empleando el procedimiento *M.dif* a partir de las estimaciones válidas otorgadas por los expertos – valores entre 1 y 12 – con lo que CALLIE-EXPERT calculó un valor real también en el intervalo [1-12] para la dificultad de cada ítem. Durante estos cálculos CALLIE-EXPERT halló 20 ítems ambiguos al utilizar intervalos contiguos de 4 niveles, a los que aplicó la definición de M.dif-2 considerando en ellos los pronósticos de 5 niveles contiguos y la desviación estándar. En concreto, los identificadores de los ítems con esta casuística fueron: 20, 35, 49, 72, 76, 82, 87, 91, 99, 107, 159, 166, 167, 171, 185, 198, 215, 222 y 231. En la Tabla 37 del anexo A4 se hallan especificados los 184 ítems con los intervalos considerados por M.dif a la hora de computar la dificultad y en la Tabla 38 del mismo anexo se detallan los cálculos realizados por CALLIE-EXPERT para obtenerla en los 20 casos de ítems con intervalos ambiguos. La aplicación de M.dif hizo descartar los juicios más extremos de la muestra final depurada que contenía 3235

entradas, de manera que realmente las estimaciones de dificultad se computaron empleando 2873 juicios.

Al finalizar la calibración CALLIE-EXPERT generó el resumen con los parámetros y resultados de la réplica para los ítems de Hezinet (Figura 52), accesible en la opción *Ver Resultados* de CALLIE-PRO. Como la comprobación del *mínimo de aportaciones por ítem* reveló que todos los cálculos de dificultad se realizaron con al menos 7 valoraciones válidas, no se marcó ningún ítem en este sentido.

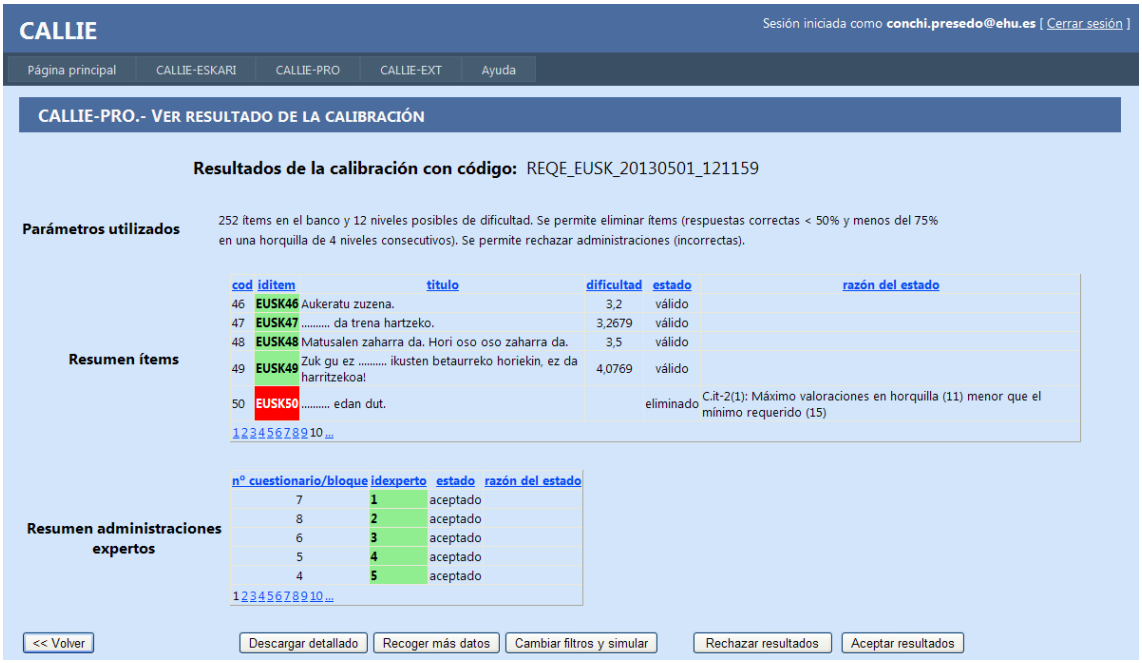


Figura 52 – Resumen de los resultados de Hezinet obtenidos por CALLIE-EXPERT.

A la vista de los resultados obtenidos por CALLIE-EXPERT, tanto en el resumen como en el informe de resultados, muchos de ellos fueron idénticos en ambos experimentos, tanto en el original como en la réplica.

En el experimento original un total de 22 ítems fueron marcados previamente como *potencialmente erróneos*. En este caso el criterio C.it-1 eliminó dos ítems anómalos (16 y 249), el criterio C.it-2 eliminó 11 ítems (59, 135,148, 170, 178, 191, 202, 225, 237, 240 y 242) y los 9 restantes permanecieron en el banco (1, 13 ,77 ,102 ,146 ,152 ,186 ,198 y 229). Con la réplica se obtuvieron estos mismos resultados.

Por otro lado, aunque la escala de dificultad del conocimiento del euskera utilizada tenía 12 niveles, prácticamente la mitad del banco de ítems tiene una dificultad estimada en alguno de los niveles 4, 5 ó 6. Concretamente en ambos experimentos, el intervalo [4, 6] concentraba el 52,1% de las estimaciones de dificultad calculadas, y la ampliación al intervalo [1,8] abarcaba a más del 90% 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. Así pues, se puede indicar que la *distribución de las dificultades estimadas* de los ítems del banco fue *idéntica* en los dos casos, de manera que la mitad del banco tiene una dificultad intermedia, y prácticamente el resto tiene dificultad media-baja, apenas habiendo ítem alguno con dificultad estimada elevada.

La única diferencia significativa en ambos experimentos fue que, frente a los 190 ítems que superaron el cribado en el experimento original, la muestra final depurada de la réplica quedó caracterizada por 184 ítems debido a que CALLIE-EXPERT aplica un redondeo más restrictivo en el criterio C.it-2.

XI 2.2 Réplica de la calibración de los ítems de IRALE

Mediante el experimento 1, que estuvo basado en los datos proporcionados por los miembros del grupo IRALE durante la calibración manual de su banco de ítems, se creó un tema denominado IRALE en CALLIE-EXPERT con 132 ítems, ficticios por motivos de confidencialidad, pero de características idénticas a los 132 ítems del banco real preparado por el grupo IRALE. Asímismo, se generó un curso en CALLIE-MOODLE para la fase de administración de estos ítems a los expertos, en el que se almacenaron los datos de los expertos y la muestra de datos recogida y revisada manualmente por los responsables de la calibración.

Para replicar la calibración ya realizada, se hizo la especificación de la calibración de los ítems aplicando las mismas restricciones que se impusieron en la calibración original, esto es, valores que fijaron los responsables de IRALE.

Al preparar el experimento 2, mediante el componente ESKARI de la herramienta se diseñó el experimento en CALLIE-EXPERT y se generó una petición de calibración utilizando el menú de cuatro pasos de CALLIE-ESKARI, con el código REQE_IRALE_20130820_112134, asociada al curso de CALLIE-MOODLE y las características iniciales que se muestran en la Tabla 22.

PASO#	CARACTERÍSTICA DE DISEÑO	VALOR INICIAL PARA IRALE
Paso 1	Conjunto de ítems a calibrar	IRALE
Paso 2	Tipo de calibración a llevar a cabo	Expertos
Paso 3	Tiempo máximo por cuestionario (minutos)	45
	Tiempo medio por ítem y parámetro (segundos)	60
	Tipo de reparto inicial de los ítems	Continuo
	Administración de ítems mediante CALLIE-MOODLE	Sí
	Datos de acceso experto	expXXX
	Tamaño homogéneo de ítems en cada cuestionario	No
	Número de cuestionarios diferentes	4
	Número de ítems por cuestionario	132, 85, 85, 67
	Número de bloques	5
	Número de ítems por bloque	47, 12, 26, 6, 41
	Número de expertos disponibles	8
	Porcentaje estimado de abandono expertos (0-100)	0
Paso 4	Método para estimar la dificultad	M.dif
	Num. niveles de dificultad (mínimo 2)	11
	Permitir dejar el nivel en blanco	Sí
	Pedir la respuesta correcta	Sí
	Obligar a dar la respuesta correcta	No
	Permitir aportaciones propias	Sí
	Permitir eliminar ítems	Sí
	% respuestas correctas para mantener el ítem (0-100)	70
	Núm. de niveles en horquilla (de 1 a número de niveles)	3

PASO#	CARACTERÍSTICA DE DISEÑO	VALOR INICIAL PARA IRALE
	% de opiniones de nivel en la horquilla (0-100)	85
	Permitir eliminar administraciones	No
	Eliminar adms. con respuesta incorrecta	No
	Qué hacer con las administraciones incompletas	Aceptar

Tabla 22 – Características iniciales de diseño para la réplica de IRALE.

Durante la introducción de estos valores iniciales, CALLIE-EXPERT generó automáticamente dos alertas debidas al bajo número de expertos captados y a la posibilidad de no llegar a las 7 opiniones por ítem. El sistema también emitió dos avisos durante la introducción de la composición de los distintos bloques y cuestionarios referidos esta vez a la posible duración excesiva de los cuestionarios y a la posibilidad de no llegar a las 7 opiniones por ítem. Ninguna de estas alertas impidió finalizar y confirmar la petición de calibración.

Los datos reales se introdujeron en la plataforma CALLIE-MOODLE para la réplica, con el objetivo de comprobar el proceso completo. Posteriormente, con la opción de CALLIE-PRO *Recoger datos desde el Moodle de CALLIE* se alimentó al sistema con los datos proporcionados por los expertos de IRALE en el curso de CALLIE-MOODLE. La transcripción de estos datos generó una muestra inicial en CALLIE-EXPERT caracterizada por 714 aportaciones (49 de ellas marcadas como eliminadas) sobre los 131 ítems no eliminados del banco y 8 expertos.

Después, mediante la opción *Depurar y Calibrar* de CALLIE-PRO, se realizó el análisis de los datos de la muestra y la calibración inicial de los ítems utilizando el workflow CA generado por la definición del experimento. Concretamente, para el análisis de datos el workflow CA se limitó al análisis de aportaciones aplicando *C.ex-1* y después ha depuró la muestra resultante utilizando solamente los criterios correspondientes a la fiabilidad del ítem, es decir, se aplicó *C.it-1* y *C.it-2*, en ese orden y solamente una vez. A continuación calibraría en dificultad los ítems no descartados aplicando M.dif. Después, con la opción *Descargar detallado*, CALLIE-EXPERT generó adecuadamente el fichero Excel con las cuatro hojas del informe de resultados correspondientes a los valores iniciales, cuyos datos más relevantes se han transcrito en las primeras tablas del anexo A5. El estudio de estas hojas ha permitido ver la evolución del proceso durante este análisis y calibración inicial, que fue idéntico al del proceso original.

La Tabla 23 ilustra la evolución que sufrió la muestra de datos a lo largo del análisis inicial. En esta tabla, *m* es el número de entradas/aportaciones de la muestra, *n* el número de ítems y *e* el número de expertos que se mantuvieron después de aplicar cada uno de los filtros considerados.

Filtro	Descripción	Aportaciones eliminadas	Ítems eliminados	Expertos eliminados	Muestra resultante
#1	Recogida datos	49	1	0	m=665, n=131, e=8
#2	C.ex-1	0	0	0	m=665, n=131, e=8
#3	C.it-1	0	0	0	m=665, n=131, e=8
#4	C.it-2	245	51	0	m=420, n=80, e=8

Tabla 23 – Evolución de la muestra durante el análisis inicial de CALLIE para IRALE.

En la recogida de datos se eliminaron las 49 aportaciones marcadas y el ítem descartado por los responsables de IRALE, pues CALLIE-EXPERT lo consideró un ítem no intentado. De este modo, la muestra para alimentar al workflow CA se redujo a 665 aportaciones válidas correspondientes a los 8 expertos y sobre un conjunto de 131 ítems. En cuanto a estos 131 ítems restantes, como más del 75% de las respuestas recogidas fueron correctas en todos ellos, no se eliminó ningún ítem mediante el filtro C.it-1. A continuación, se les aplicó C.it-2 que eliminó 51 ítems más junto con las 245 aportaciones correspondientes, por no alcanzar el 85% de opiniones dentro de la horquilla establecida. Aunque el experto "exp8" solo obtuvo una tasa de acierto del 65,88% de respuestas correctas su administración no se rechazó, puesto que se había definido la no aplicación de C.ex-2. Por tanto, la muestra de partida se redujo a 420 aportaciones válidas correspondientes a 8 expertos y 80 ítems.

Finalizado el análisis inicial de los datos, el workflow CA llevó a cabo la calibración inicial de los ítems no descartados y obtuvo los 80 niveles de dificultad correspondientes empleando el procedimiento *M.dif* a partir de las estimaciones válidas otorgadas por los expertos – valores entre 1 y 11 – con lo que CALLIE-EXPERT calculó un valor real también en el intervalo [1-11] para cada ítem. Durante estos cálculos se puso de manifiesto la existencia de 1 ítem (el ítem 109) con la casuística de estar comprendido en dos intervalos contiguos de 3 niveles con la misma tasa de frecuencias de pronósticos dificultad. Por tanto, resultó ser ambiguo en M.dif-1 y se le aplicó M.dif-2. La aplicación de M.dif eliminó 9 aportaciones más, con lo que solamente se consideraron 411 aportaciones que es el 57,56% de las recogidas.

Para obtener los resultados de calibración definitivos para IRALE, mediante la opción de CALLIE-PRO *Cambiar filtros y simular* se relajó el porcentaje de opiniones de nivel en la horquilla hasta el 60%, del mismo modo que se había hecho en el experimento original. De nuevo, pero esta vez con la opción *Simular y Descargar en Excel*, CALLIE-EXPERT generó adecuadamente el fichero Excel con las cuatro hojas del informe de resultados correspondientes a este segundo caso, cuyos datos más relevantes se han transcrito en las dos últimas tablas del anexo A5. Al igual que en el caso inicial, los resultados fueron idénticos en ambos experimentos, original y réplica.

Así, en cuanto a la evolución del proceso en este segundo caso, CALLIE-EXPERT rehizo los cálculos con la misma muestra de datos que en el caso anterior, cuyo tamaño inicial era de 714 aportaciones. Una vez eliminado el ítem 125, sus 2 aportaciones junto con los juicios eliminados por los expertos principales la muestra contaba con 665 aportaciones. En este caso el análisis de dispersión (C.it-2) no eliminó ningún ítem y se aplicó el procedimiento M.dif a los 131 ítems no descartados. Durante estos cálculos 12 de los ítems resultaron ser ambiguos en M.dif-1 – concretamente los ítems 14, 17, 36, 42, 46, 56, 62, 71, 75, 95, 100 y 109 – y se les aplicó M.dif-2 a una horquilla de 4 niveles. La aplicación de M.dif supuso la eliminación de 18 aportaciones más. Por tanto, el workflow CA utilizó 647 aportaciones – que es el 90,62% de las recogidas – para calcular la dificultad de los 131 ítems del banco. Más del 80% de los ítems resultaron ser de dificultad baja o media y solo uno de dificultad muy alta. También destaca que no hubo ningún ítem valorado con la dificultad máxima (C2). De hecho, la mayor dificultad obtenida fue de 9,67 en la escala [1-11].

Al finalizar la calibración CALLIE-EXPERT generó el resumen con los parámetros y resultados de la réplica para los ítems de IRALE (Figura 53), accesible en la opción *Ver Resultados* de CALLIE-PRO. Como la comprobación del *mínimo de*

aportaciones por ítem reveló que solamente en 15 de los 131 ítems que permanecían en el banco – los ítems 25, 28, 29, 30, 33, 34, 35, 40,51, 53, 74, 84, 88, 89 y 109 – los cálculos de dificultad se realizaron con al menos 7 valoraciones válidas. Todos los demás ítems fueron etiquetados y se visualizaron en naranja junto al aviso correspondiente, como aparece en el resumen de ítems.

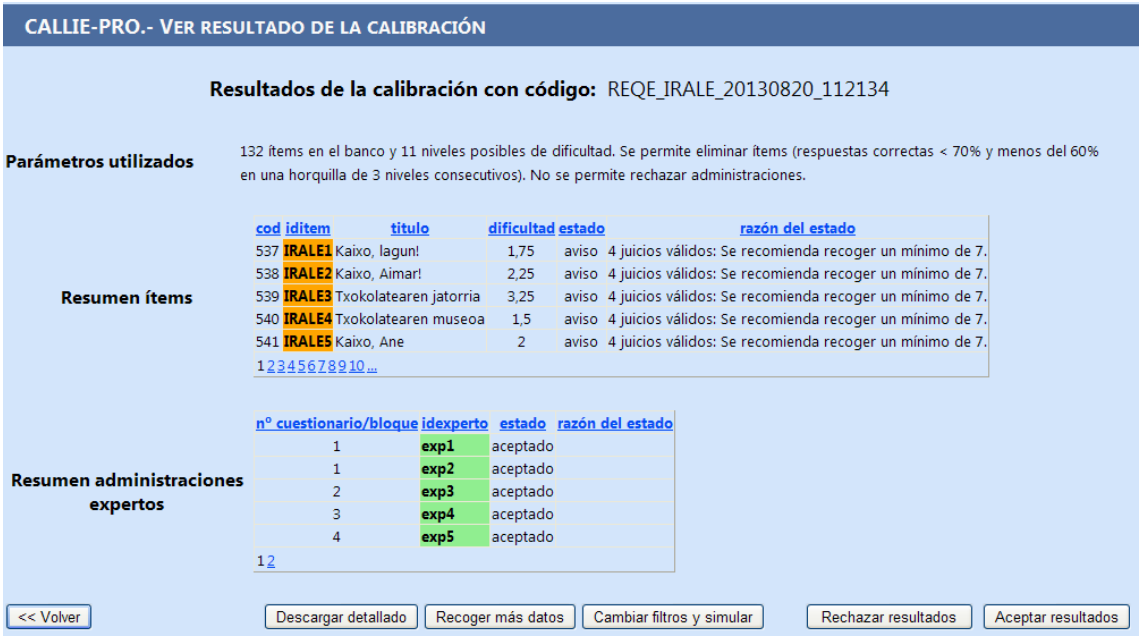


Figura 53 – Resumen de los resultados de IRALE obtenidos por CALLIE-EXPERT.

XI 3 Pruebas globales para la aplicación Web

Las pruebas relevantes para esta memoria han abarcado la recopilación de impresiones generales sobre la herramienta CALLIE a través de la evaluación del módulo EXT, mediante una encuesta similar a la que se había hecho en los experimentos 1 y 2, en la que los usuarios otorgaron puntuaciones de 1 a 10 a distintas características de la aplicación Web. En Irastorza (2014) se pueden encontrar todos los detalles sobre las pruebas realizadas.

Para la interpretación de los resultados obtenidos en estas pruebas cabe destacar que, antes de la creación del módulo EXT, se había mejorado la utilidad ya existente en CALLIE-EXPERT para la recogida de datos desde hojas de Excel, en vista de los pobres resultados obtenidos en las pruebas ya realizadas al componente con la interfaz original.

En la Tabla 24 se recogen los resultados que se obtuvieron por grupos con relación a la aplicación Web, donde se estudiaron la facilidad de uso, la idoneidad en el formato de las interfaces, la velocidad de respuesta y el interés de la herramienta. Al Grupo1 pertenecen usuarios inexpertos que usan el ordenador para ocio y ofimática (2 amas de casa y 1 jubilado), al Grupo2 usuarios que están o han estado indirectamente en contacto con herramientas o funcionalidades que se utilizan en el módulo EXT (3

graduados en ingeniería informática) y al Grupo3 usuarios potenciales de CALLIE (3 profesores universitarios).

	Grupo1	Grupo2	Grupo3
Facilidad de uso	6	9	8
Formato de las interfaces	8	8	8
Velocidad de respuesta	9	8,5	8
Interés	8	8,5	8

Tabla 24 – Resultados por grupos para la aplicación Web.

En cuanto a las opiniones de los participantes sobre las funcionalidades de CALLIE-EXT de la aplicación Web, idénticas en interfaz al resto de CALLIE-EXPERT, juzgaron que el formato de su interfaz era adecuado y además que era fácil de usar (excepto en el Grupo1). La idea general del sistema CALLIE les pareció interesante y juzgaron que las nuevas funcionalidades creadas son de utilidad para CALLIE y que funcionan correctamente.

Los resultados globales que se obtuvieron para la aplicación Web se muestran en la Figura 54, esto es, una media de 7,67 como puntuación en facilidad de uso, un 8 en el formato de las interfaces, un 8,5 en velocidad de respuesta y un 8,25 en interés de la herramienta.

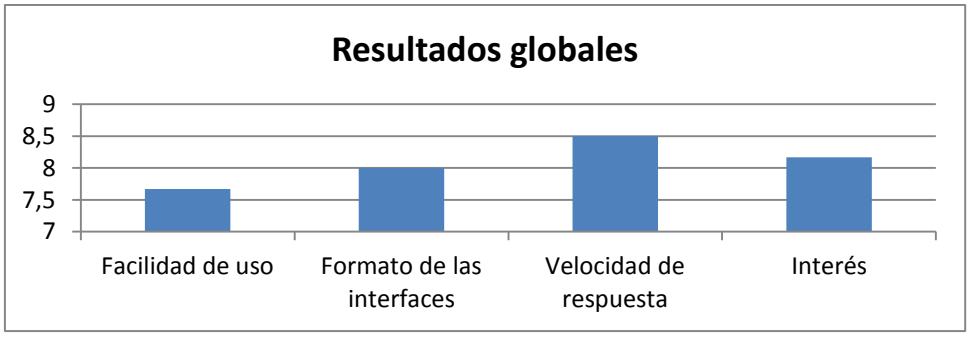


Figura 54 – Resultados globales para la aplicación Web.

XI 4 Pruebas de integración con CALLIE-TRI

Con estas pruebas se ha verificado que una vez alcanzado el punto de entrada al proceso de calibración correspondiente – vía expertos o psicométrico – CALLIE sigue los pasos correspondientes al tipo de proceso elegido, solicitando los datos apropiados y almacenando las especificaciones seleccionadas adecuadamente.

Para integrar sus dos componentes CALLIE-EXPERT y CALLIE-TRI, la herramienta CALLIE utiliza la interfaz del menú de cuatro pasos de ESKARI, en el que el punto de entrada se corresponde con el paso 2 de ESKARI, y el de salida con la opción de envío/aceptación de la petición, con la que CALLIE genera el modelo de calibración correspondiente y lo almacena en el sistema. Tanto el menú de ESKARI como la base de datos, que utiliza CALLIE para guardar los modelos generados, son

comunes a ambos tipos de calibración. Así, la validación de la integración, tiene que ver con la navegación por este menú de la herramienta y con la integridad de la base de datos común.

Al ser sobre todo un tema técnico, en la realización de estas pruebas no se utilizaron usuarios, sino que ha sido la propia autora de esta memoria la que ha verificado esta integridad. Previamente, para posibilitar esta verificación, la autora de esta memoria ha creado un script en el estándar SQL (Structured Query Language) que genera automáticamente tanto el modelo de datos común conforme a los metamodelos de ambos tipos de calibración, una base de datos llamada CALLIE_DB, como el banco de ítems de prueba EUSK dentro del modelo. Una vez generado este modelo de datos se ha comprobado satisfactoriamente, utilizando el banco de ítems de prueba EUSK, que una vez elegido un tipo de calibración éste no se puede cambiar a no ser que se anule la petición, y que navegando por el resto de pasos del menú, CALLIE va generando el modelo de calibración en XML correspondiente al tipo de calibración: vía expertos o psicométrica. También se ha verificado que, una vez aceptada la petición, este modelo se guarda correctamente en la base de datos, en la que además se comprueba automáticamente la completitud del modelo XML mediante el esquema XSD correspondiente al tipo de petición.

XII Conclusiones

Este capítulo concluye la exposición del trabajo realizado para la culminación del presente proyecto. En primer lugar, se resumen las aportaciones realizadas por el desarrollo del sistema CALLIE para la calibración de bancos de ítems mediante el juicio de expertos (sección XII 1). A continuación, se describen las posibles vías de investigación y desarrollo que quedan abiertas y que permitirán al actual sistema evolucionar y mejorar su funcionalidad (sección XII 2). Para finalizar, se enumeran las publicaciones de la autora que han visto la luz durante el transcurso del trabajo realizado en el presente proyecto (sección XII 3).

XII 1 Principales aportaciones

Las principales aportaciones de esta investigación surgen al dar cobertura a los distintos objetivos que se plantearon al inicio de la misma (sección I 2) y así se han ordenado y clasificado en los párrafos que conforman el resto de este apartado.

Para garantizar la integración (*objetivo 1*) se ha analizado el proceso global de calibración de un banco de ítems independientemente del método seguido y se ha obtenido el conocimiento necesario para establecer una base común sobre la que poder realizar el diseño, control y seguimiento de cualquier proceso de calibración. Con este conocimiento, *se ha diseñado e implementado el sistema de calibración CALLIE*. El sistema utiliza técnicas de desarrollo software dirigido por modelos, concretamente la generación automática de software, y se basa en la tecnología de workflows para la creación de actividades que se pueden organizar de diferentes modos. Esta organización en workflows permite que se puedan añadir de manera sencilla nuevas actividades y que se puedan reorganizar las existentes. Gracias a este tipo de diseño, también se puede comprobar el resultado de hacer cálculos alternativos a una tarea. La implementación de este sistema de calibración se ha plasmado en una herramienta informática de ayuda: CALLIE. La *herramienta CALLIE* es una aplicación Web que, utilizando tecnología .NET, automatiza el proceso global de calibración de un banco de ítems y ofrece al usuario la posibilidad de utilizar para ello cualquiera de los dos métodos más usuales: calibración mediante expertos y calibración psicométrica. En concreto, la herramienta permite a los usuarios registrados calibrar los bancos de ítems que ellos mismos podrán introducir o seleccionarlos de un repositorio en el que los ítems se encuentran en formato IMS QTI. Se supone que estos usuarios serán principalmente profesores, aunque su uso no está restringido especialmente a este colectivo, sino a cualquiera que desee calibrar un banco de ítems y que – en el caso de calibraciones vía expertos – utilice una escala finita de valores sucesivos de calibración para que cada experto estime el nivel de dificultad, esto es, una escala transformable en números enteros positivos.

Para diseñar e implementar el modelo de especificación (*objetivo 2*) se ha analizado el proceso de calibración de un banco de ítems mediante el juicio de expertos y se ha obtenido el conocimiento necesario para poder especificar el experimento de administración de ítems que conlleva, concretado en las decisiones que se tienen que tomar y cómo interfieren entre ellas. De este conocimiento *ha surgido ESKARI*, el componente de especificación del experimento de calibración. Se ha generado un *modelo de calibración vía expertos* en XML para CALLIE-EXPERT, que ha sido unificado con el modelo correspondiente a la calibración psicométrica, generado en CALLIE-TRI, para garantizar la integrabilidad de los dos métodos en un único sistema.

Durante el diseño e implementación del subsistema de administración de ítems (*objetivo 3*) es cuando han surgido – a juicio de la autora – las mayores aportaciones de este trabajo. En primer lugar se ha llevado a cabo un *estudio de los estándares IMS* relacionados (capítulo IV) y de las distintas *plataformas educativas* que soportan la importación de paquetes IMS (sección VII 1) y de entre ellas se han estudiado en profundidad las características de la plataforma Moodle para cubrir la administración de cuestionarios a los expertos utilizando dicho estándar. En segundo lugar, se han estudiado los métodos de aplicación de un cuestionario a un experto y *se ha creado ADMINQ Factory*, un módulo que, utilizando las técnicas de desarrollo dirigido por modelos y una plataforma educativa preexistente que permita la importación de paquetes IMS, genera automáticamente un sistema de administración de cuestionarios a expertos basado en web. Para ello *se ha propuesto un nuevo modelo de ítem* a utilizar en las calibraciones mediante expertos, que surge de los ítems de evaluación más comunes. Este modelo de ítem ha sido implementado en IMS QTI lo que posibilita su importación directa a cualquier sistema que contemple ese estándar con independencia de la versión utilizada. En virtud de este modelo CALLIE-EXPERT, en los cuestionarios a rellenar por cada experto, siempre genera tres campos para cada ítem a calibrar: el primero corresponde al propio ítem que el experto deberá calibrar, en el segundo campo el experto estimará el nivel de dificultad para ese ítem mediante un valor adecuado a la escala adoptada, y en el tercero podrá especificar cualquier otro dato sobre el ítem si así se lo indica el responsable de la calibración. Partiendo de este nuevo modelo de ítem y de la especificación de la calibración *se ha creado un programa software capaz de obtener el paquete IMS CC* con el que posteriormente se genera un curso en Moodle para administrar los cuestionarios a los expertos disponibles. Asímismo, *se ha ideado un servicio Web* para Moodle que es capaz de importar paquetes IMS CC creando automáticamente el curso de administración y *se han detectado servicios ya existentes* en la plataforma para el alta y la matriculación automática de expertos en el curso previamente generado. Por último, se ha preparado el modelo de datos para su soporte.

La propuesta llevada a cabo en esta investigación, y detallada en el párrafo anterior, implica que el sistema de calibración así definido puede aprovechar todas las funcionalidades que posea la plataforma educativa en la que se crea y, además, abre una vía que permite la integración de los procesos de evaluación y calibración en una única herramienta software. Por su parte, el uso de estándares e-learning – en este caso los de IMS, que es una de las iniciativas más populares con más de 20 estándares disponibles de forma gratuita en Internet – proporciona múltiples ventajas, entre ellas permite la creación de contenidos reutilizables y facilita la interoperabilidad entre distintos sistemas. Además, los ítems a calibrar no tienen por qué ser de selección múltiple, ya que al estar el repositorio representado en el formato IMS QTI los ítems podrían corresponder a cualquier otro tipo soportado por el estándar, aunque la herramienta no

soporta su introducción en la actualidad. Por último, la autora considera acertada la elección de Moodle como plataforma Web para gestionar las administraciones de los ítems a los expertos, puesto que es muy popular en el ámbito educativo y los usuarios potenciales de la herramienta CALLIE-EXPERT pueden estar ya familiarizados con ella, al menos en el rol de profesor, es gratuita, y al ser de código abierto permite añadir nuevas funcionalidades.

El diseño e implementación del subsistema de análisis y cálculos (*objetivo 4*) también ha proporcionado varias aportaciones destacables. En primer lugar, se han estudiado los diferentes análisis de administraciones que se pueden llevar a cabo para poder asegurar la calidad de las aportaciones de los expertos y *se han creado análisis de administración* susceptibles de ser incluidos en un workflow para hacer una revisión automática de las administraciones realizadas. Asímismo, se han analizado distintos estudios y cálculos que se pueden efectuar tanto a los ítems como a las administraciones y después estos *estudios y cálculos se han organizado* en forma de flujos de trabajo a ejecutar sobre los datos recopilados para poder sistematizar su aplicación. Con este conocimiento se ha creado una biblioteca que contiene los distintos elementos necesarios para la generación de los workflows detectados. En segundo lugar, *se han utilizado técnicas de desarrollo dirigido por modelos* para crear el código correspondiente a los flujos de trabajo que se tienen que ejecutar como consecuencia del modelo de especificación generado por ESKARI. Esta propuesta se ha plasmado en el módulo *WF Factory* de CALLIE-EXPERT, que genera dos workflows distintos para realizar la calibración de los ítems. El sistema de generación de workflows solo tiene que tener conocimiento de los elementos disponibles, para generar la correspondiente actividad en el workflow. Además, el sistema crea contenedores de datos adecuados a los resultados que se van a obtener para cada estudio que se especifica, siempre de acuerdo con el metamodelo de los resultados que se obtienen. En tercer lugar, *se han creado actividades de workflow para recoger los datos*, que pueden ser incluidas en la especificación del estudio a realizar sobre administraciones. Estas actividades son miniprogramas que pueden interactuar con otros sistemas externos como Moodle y MS Excel con objeto de alimentar los workflows con los datos aportados por los expertos. Mediante este tipo de actividades se consigue la integración de este subsistema con el subsistema de administración. En cuarto lugar, *se han analizado los cálculos detallados* llevados a cabo durante el proceso de calibración y se han estudiado distintos *programas software* (sección VIII 1) que permiten realizar los cálculos necesarios en una calibración, tanto a nivel de depuración de los datos recogidos de los expertos como a nivel del propio cálculo del parámetro dificultad a partir de esos datos. Con este conocimiento se ha enriquecido el metamodelo de resultados con los datos correspondientes a esos cálculos. Por último, basadas en los modelos de datos previamente diseñados, *se han creado actividades de workflow para generar automáticamente una serie de informes* en formato MS Excel con las especificaciones del experimento, los datos aportados por los expertos, los cálculos detallados y los resultados concernientes al proceso de calibración.

En cuanto al diseño e implementación de la herramienta (*objetivo 5*) se ha creado una interfaz informal, sencilla y flexible, que dispone de varios tipos de ayuda para guiar al responsable a lo largo de todo el proceso de calibración, sin que sea necesario que éste cuente con conocimientos previos en el área (las características concretas se han descrito en el capítulo X de esta memoria). Se ha creado el *módulo PRO-C* que controla el avance de cada proceso de calibración siguiendo las tareas que

se tienen que ejecutar como consecuencia del modelo de especificación generado por ESKARI. A través de la interfaz de este módulo, denominada *CALLIE-PRO*, el responsable puede controlar a su vez sus propias calibraciones. Además, esta interfaz posibilita la ejecución de ciertas tareas de manera externa a la herramienta, por ejemplo, el responsable puede utilizar otros sistemas de administración de ítems y puede llevar a cabo otros métodos de análisis y cálculo – tanto alternativos como complementarios – a los realizados por CALLIE-EXPERT. Así, para dar cobertura a procesos de calibración cuyas administraciones no han sido realizadas con la plataforma Moodle integrada en el sistema, la herramienta es capaz de capturar los datos de las administraciones a expertos desde un fichero MS Excel con un formato preestablecido. Por otra parte, para estudiar y revisar el proceso, el usuario tiene acceso a los *informes predefinidos generados por el subsistema de análisis y cálculos*. Como estos informes se descargan en formato MS Excel, la información que contienen puede ser objeto de nuevos estudios con otros programas software que soporten la importación de datos desde Excel, como son la mayoría de las herramientas mencionadas en la sección VIII 1, por ejemplo R.

Para probar la validez del sistema (*objetivo 6*) se han comparado los resultados obtenidos por la herramienta CALLIE-EXPERT con los conseguidos para la calibración mediante el juicio de expertos en dos casos reales: el que ya fue utilizado para la generación de TAIs de Hezinet y el caso IRALE, una calibración externa asesorada por miembros del grupo de investigación GHyM. Además se han realizado una serie de pruebas para verificar que cada componente de CALLIE-EXPERT cumple su función y que la integración con CALLIE-TRI se realiza apropiadamente.

Como *aportaciones adicionales* cabe destacar que se ha creado un *banco de ítems de prueba* con el que el usuario puede efectuar simulaciones de calibración con distinta cantidad de ítems, tanto basadas en el juicio de expertos como fundamentadas en la TRI. Asímismo, en la calibración mediante el juicio de expertos se han aprovechado los workflows que crea WF Factory, para implementar un módulo que permite *simular resultados alternativos* que se obtendrían al aplicar distintos estudios a los mismos datos recogidos. En último lugar, *se ha creado un script en el estándar SQL* (Structured Query Language) que genera automáticamente el modelo de datos común, CALLIE_DB, lo que permite que la herramienta pueda adaptarse fácilmente a distintos sistemas de gestión de bases de datos y facilita la portabilidad del sistema a distintos entornos software. Este script también genera el banco de ítems de prueba.

XII 2 Principales líneas futuras de trabajo

La presente memoria ha descrito el trabajo llevado a cabo durante el transcurso de esta tesis, pero se podrían haber tomado otras vías de desarrollo tanto para el sistema ideado como para la formalización de la calibración de bancos de ítems en sí. Seguidamente se presentan una serie de líneas abiertas para mejorar y dar continuidad a la labor realizada.

La primera línea futura consiste en *mejorar la herramienta solucionando las debilidades detectadas* por la autora de esta memoria, a saber: actualmente varias decisiones se toman basándose en criterios fijos que son consecuencia de la experiencia,

los métodos implementados para realizar automáticamente el diseño de cuestionarios de los bancos de ítems son rudimentarios, CALLIE-EXPERT solo permite la introducción de ítems individuales de respuesta múltiple, con un enunciado y varias posibles respuestas textuales, y el sistema implementa exclusivamente los análisis de administraciones, estudios y método de calibración que utilizó Arruabarrena (2010). Con objeto de mejorar la herramienta, se podrían revisar todas las decisiones de ESKARI y parametrizar al máximo los procesos de decisión; mejorar los métodos de diseño de cuestionarios incluyendo una serie de criterios más avanzados para realizar el reparto de los ítems; ampliar las posibilidades de calibración de ítems soportando otros formatos de ítems en el sistema que ya contempla el estándar IMS QTI, como localizar un punto en un dibujo o foto, unir puntos para formar una figura, presentación de videos o música, o una sopa de letras o incluyendo nuevos métodos de especificación de los bancos de ítems; e incluso se podrían ampliar las variantes de actividades para integrar en el workflow CA. Con esta última propuesta se enriquecería el sistema con métodos alternativos a M.dif para estimar la dificultad, con nuevas propuestas para el análisis de administraciones e ítems, o añadiendo otro tipo de estudios.

Más trabajo futuro en esta primera línea consiste en *mejorar la herramienta subsanando las carencias detectadas* por los usuarios durante las pruebas de evaluación de sus distintos componentes, que fueron: problemas en la introducción de ítems directamente en formato IMS QTI e imposibilidad de introducir varios ítems a la vez, ayuda insuficiente en cuanto al diseño de cuestionarios no homogéneos, ayuda insuficiente de la plataforma Moodle especialmente en cuanto a la distribución y envío de cuestionarios, e imposibilidad de ordenar la lista de peticiones. Para solventar las carencias detectadas en el componente ESKARI se podría añadir una utilidad que permita subir a CALLIE varios ítems en IMS QTI simultáneamente, y completar la ayuda de CALLIE-EXPERT para la creación de cuestionarios incluyendo definiciones más exhaustivas con ejemplos de los conceptos utilizados por la herramienta. En cuanto a las carencias detectadas en los demás componentes, en CALLIE-MOODLE se podría completar la ayuda que ofrece Moodle sobre el funcionamiento de los cuestionarios, tanto para el rol de profesor como para el rol de estudiante, y en la interfaz de CALLIE-PRO se podría posibilitar la ordenación por varios campos de la lista de peticiones, al igual que ya sucede con las listas de resultados que muestra este mismo componente.

Una segunda vía de mejora consiste en *automatizar aún más la herramienta*. En primer lugar, se podría ampliar CALLIE-EXPERT para que el sistema fuese capaz de *crear informes de forma automática conocidas las características* que se quieren analizar. También podría interesar la generación de estadísticas demográficas e informes sobre los expertos que han participado en la calibración. Por otro lado, aunque el responsable puede gestionar a los expertos y controlar su participación a través de Moodle de forma manual, se podría *añadir un nuevo workflow para la gestión de captación de participantes*. Esta línea consiste en especificar e implementar un nuevo workflow que sea capaz de controlar automáticamente desde CALLIE-EXPERT si el experto envía las respuestas, que confirme su participación, etc. Por último, en el caso de un sistema combinado que utilice cuestionarios en papel, CALLIE-EXPERT permite la importación de los datos de las administraciones en formato Excel. Sin embargo, sería útil la inclusión de un *módulo que permita incorporar las administraciones que se han hecho manualmente*, empleando aparatos de lectura óptica para capturar las respuestas de forma automática.

Una tercera vía de investigación tiene que ver con los *dispositivos y las plataformas utilizadas en la implementación de la herramienta*. En este sentido, se ha trabajado en la creación de un *módulo para smartphones* que facilita la gestión de expertos en las administraciones. Esta versión para móviles permite al responsable comunicarse con Moodle creando una cuenta para un nuevo experto, matriculándolo en el curso que se le indique como alumno y/o eliminando a un experto previamente creado, utilizando los servicios web del propio Moodle. En un futuro se podría completar este módulo de forma *que soporte todas las funcionalidades de CALLIE-EXPERT*, pero a través del dispositivo móvil. Por otra parte, el programa software que utiliza CALLIE-EXPERT para generar cartuchos IMS CC con el curso de administración para calibración vía expertos también está preparado para generar paquetes de contenido IMS CP (Smythe y Nielsen, 2007) con el curso, que es otro método de distribuir contenidos entre sistemas. Se podría explorar la posibilidad de administrar los ítems a los expertos mediante el *uso de otras plataformas educativas diferentes a Moodle* que admitan cualquiera de estos dos estándares, como por ejemplo Atutor, Claroline, Ilias, Sakai o cualquier otra de las mentadas en la sección VII 1.

Una cuarta línea es la relativa a las *teorías utilizadas tanto en la preparación del banco de ítems como en su posterior calibración*. Por un lado, la generación e inclusión de ítems en el sistema es un trabajo costoso, tanto en tiempo como en esfuerzo, por lo que se podría contemplar la generación automática de ítems. Esta vía de investigación consistiría en *incluir la especificación de ítems parametrizados para su generación automática.* Los denominados modelos generativos de ítems se han definido para captar el conocimiento de los expertos que se invierte en la creación de bancos de ítems, de manera que se pueda automatizar no sólo la composición de nuevos ítems de evaluación, en tiempo real incluso, sino además controlar a priori la dificultad de las nuevas preguntas generadas (Rojas Tejada, 2001). Por otro lado, también se podría considerar la *aplicación de otras teorías de calibración*. El sistema se basa en la actualidad en ítems independientes, sin embargo, en algunas ocasiones es necesario que algunos ítems se presenten al usuario en cierto orden o que vayan ligados entre sí. Para modelar este tipo de ítems o conjuntos de ítems existe la teoría de respuesta al testlet (Wainer, Bradlow y Du, 2000), que no se contempla en CALLIE. Es una línea abierta que puede tener muchas aplicaciones, una de ellas es la evaluación del comportamiento en casos clínicos para estudiantes de medicina, en la que se está trabajando actualmente dentro del grupo Erabaki de la UPV/EHU.

Por último, se podría investigar la *aplicación del proceso de generación de workflows a otras áreas.* Hasta el momento, el módulo WF Factory se ha pensado para la generación de actividades dentro del ámbito de la calibración de ítems. Sin embargo, durante el desarrollo de la tesis, se ha descubierto que ese enfoque podría ser de utilidad a otras áreas. Por ejemplo, la generación de workflows podría aplicarse en sistemas de control de estudios clínicos o en la informatización de guías clínicas y, de hecho, ya se está haciendo una pequeña incursión en el área dentro del grupo Erabaki de la UPV/EHU (López-Cuadrado, Armendariz, Presedo, Segundo, Barrena, Korta y Pérez, 2015).

XII 3 Publicaciones

En este apartado se presentan las publicaciones de la autora que han tenido que ver con la elaboración de esta tesis directa o indirectamente. En total han sido 1 libro, 5 artículos y 7 ponencias en congresos, y a nivel internacional incluyen 2 revistas y 7 congresos. Actualmente se están preparando varios artículos basados en el trabajo descrito en esta memoria – unos centrados en la parte de calibración mediante expertos y otros en el sistema CALLIE – con objeto de buscar la aprobación de la comunidad académica, tanto nacional como internacional, acerca de las ideas y resultados que se han ido obteniendo.

Directamente relacionado y generado por esta tesis, hasta este momento se ha publicado un libro (Presedo, Armendariz y López-Cuadrado, 2012) que une los dos tipos de calibración (psicométrico y por mediación de expertos) mediante el análisis de los distintos aspectos a tener en cuenta desde el momento en que se toma la decisión de recurrir a los test informatizados como mecanismo de evaluación, hasta concluir con la calibración. En él se detalla la informatización de las distintas tareas correspondientes a ambos métodos de calibración y también se revisan distintos programas software para llevarlas a cabo. Además se han publicado tres artículos, dos en revistas internacionales, y uno en revista nacional, más una aportación a un congreso sobre el sistema CALLIE para calibraciones vía expertos:

- Concepción Presedo, Ana J. Armendariz y Javier López-Cuadrado (2012). *Calibración de ítems para test informatizados: Descripción detallada de las fases en la construcción de test de evaluación adaptativos mediante ordenador*. 76 págs. Editorial Académica Española. ISBN: 3846576492.
- Concepción Presedo, Ana J. Armendariz, Javier López-Cuadrado y Tomás A. Pérez (2015). *Calibración de ítems vía expertos utilizando Moodle*. Revista Iberoamericana de Educación. Volumen 69, Número 1, pp 117-132, ISSN: 1022-6508 / ISSNe: 1681-5653. Organización de Estados Iberoamericanos (OEI/CAEU)
- Concepción Presedo, Ana J. Armendariz, Javier López-Cuadrado y Tomás A. Pérez (2015). *Sistema de ayuda para la calibración de ítems por el procedimiento basado en el juicio de expertos*. Revista Internacional de Tecnologías en la Educación, Volumen 2, Número 1, pp 1-15, ISSN: 2386-8384
- Concepción Presedo, Ana J. Armendariz, Javier López-Cuadrado y Tomás A. Pérez (2014). *Sistema de ayuda para la calibración de ítems por el procedimiento basado en el juicio de expertos*. XXI Congreso Internacional sobre Educación y Aprendizaje, Universidad de Touro (Nueva York, USA)
- Ana J. Armendariz, Javier López-Cuadrado, Tomás A. Pérez y Concepción Presedo, (2016). *Azterketa informatizatu eraginkor baten bila*. EKAIA, Euskal Herriko Unibertsitateako Zientzi eta Teknologi Aldizkaria. DOI: 10.1387/ekaia.16368.

Por último, se ha presentado una ponencia en un congreso internacional sobre evaluación en tecnología sanitaria, referida a las guías clínicas:

- Javier López-Cuadrado, Ana J. Armendariz, Concepción Presedo, U. Segundo, R. Barrena, J. Korta y Tomás A. Pérez (2015). *Why tables on clinical practice guidelines are not easily computerizable*. HTAi 12th Annual Meeting, Oslo (Norway). 15-17 June, 2015. Health Technology Assessment international.

Indirectamente relacionados con esta tesis, se han publicado otros cuatro artículos y dos ponencias en revistas y congresos internacionales bajo el tema de la visualización de proyectos software mediante diversas metáforas y que tienen que ver con la gestión del software, la gestión del diseño y los interfaces de usuario en ingeniería del software. Todos ellos han resultado de especial utilidad a la hora de implementar la interfaz de CALLIE aplicando los principios de ingeniería del software:

- Amaia Aguirregoitia, J. Javier Dolado y Concepción Presedo (2010b). *Software Project Visualization Using Task Oriented Metaphors.* Journal of Software Engineering and Applications, JSEA 3(11): 1015-1026.
- Amaia Aguirregoitia, J. Javier Dolado y Concepción Presedo (2010c). *Using the magnet metaphor for multivariate visualization in Software management*. Visual Analytics in Software Engineering - VASE 2009 as part of the IEEE/ACM International Conference on Automated Software Engineering, Auckland, (New Zealand). Technical Report ECSTR10-11, July 2010 ISSN 1179-4259 School of Engineering and Computer Science. Pages 17-24.
- Amaia Aguirregoitia, J. Javier Dolado y Concepción Presedo (2008a). *A landscape metaphor for visualization of software projects*. SoftVis '08: Proceedings of the 4th ACM symposium on Software visualization, Herrsching am Ammersee (Germany). Pages 197-198.
- Amaia Aguirregoitia, J. Javier Dolado y Concepción Presedo (2008b). *A metro map metaphor for visualization of software projects*. SoftVis '08: Proceedings of the 4th ACM symposium on Software visualization, Herrsching am Ammersee (Germany). Pages 199-200.
- Amaia Aguirregoitia, J. Javier Dolado y Concepción Presedo (2010a). *Applying the metro map to software development management*, Proc. SPIE 7530, Visualization and Data Analysis 2010, San José, CA. (USA).
- Amaia Aguirregoitia, J. Javier Dolado y Concepción Presedo (2009). *Software Visualization Using a Treemap-hypercube Metaphor*. International Conference on DMS – Distributed Multimedia Systems 2009, San Francisco (USA).

Además, se ha participado en tres ponencias en congresos nacionales sobre métricas a utilizar en la coordinación de proyectos software basándose en objetivos y modelos de madurez de ingeniería del software y que han resultado de utilidad para gestionar el progreso y la puesta en común de la herramienta CALLIE:

- Concepción Presedo, J. Javier Dolado y Amaia Aguirregoitia (2010). *Estudio de métricas para el control de proyectos software.* Actas del 10° Taller de las Jornadas sobre apoyo a la decisión en Ingeniería del Software y Bases de Datos, 4(1):65-72.
- Concepción Presedo y J. Javier Dolado (2007). *Medición Práctica de la Coordinación utilizando GQ(I)M y CMMi*. Actas del 8º Taller sobre Apoyo a la Decisión en Ingeniería del Software. II Congreso Español de Informática, Zaragoza (España).
- Concepción Presedo y J. Javier Dolado (2006). *El problema de la coordinación en proyectos software: Enfoque mediante Sistemas Multiagente.* XI Jornadas de Ingeniería del Software y Bases de Datos, Sitges (España).

Parte Quinta: Anexos y Bibliografía

La Parte Quinta está dedicada a los **anexos y bibliografía**, y se divide de la siguiente manera:

El anexo **A1 - Ejemplo de modelo de calibración** recoge un ejemplo de modelo de calibración vía expertos en XML conforme al MMCA.

El anexo **A2 - Experimento 1: Calibración de los ítems del tema IRALE** incluye la génesis, los detalles de las pruebas y los resultados relevantes del experimento 1 que ha permitido calibrar vía expertos con CALLIE-EXPERT un banco de ítems, que había sido calibrado previamente de forma manual en colaboración con miembros de IRALE, y también evaluar distintas funcionalidades de CALLIE-EXPERT.

El anexo **A3 - Experimento 2: Réplicas y evaluación con alumnos** contiene los detalles de las pruebas y los resultados relevantes del experimento 2. Este segundo experimento complementa la evaluación de las distintas funcionalidades de CALLIE-EXPERT y permite replicar las calibraciones de los bancos de ítems llevadas a cabo manualmente para Hezinet e IRALE.

El anexo **A4 - Informe de resultados réplica Hezinet** contiene los resultados obtenidos por CALLIE-EXPERT en la réplica del proceso de calibración vía expertos del banco de ítems de Hezinet.

El anexo **A5 - Informe de resultados réplica IRALE** presenta en detalle los resultados obtenidos por CALLIE-EXPERT en la réplica del proceso de calibración vía expertos del banco de ítems de IRALE.

El último apartado está dedicado a las **Referencias bibliográficas** y recoge la bibliografía utilizada y referida, con el objetivo de que el lector o lectora del presente documento pueda profundizar en los aspectos tratados.

A1 Ejemplo de modelo de calibración

Las especificaciones del metamodelo MMCA se almacenan mediante un *modelo de calibración en XML* formado por tres partes diferenciadas: datos generales, detalles para los formularios y filtrado. Cada una de estas partes existirá o no dependiendo de si el responsable de la calibración ya ha tomado esa decisión o no. Existe un único esquema XSD que define la estructura del modelo XML con independencia del tipo de calibración – mediante expertos o psicométrico – y contra el que se validan todos estos modelos antes de su almacenamiento. En la Figura 55 se detalla un ejemplo para calibración vía expertos.

```
1  <?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
2  <request>
3    <code>REQE_EUSK12_20150210_091706</code>
4    <user>responsible.of.calibration@ehu.es</user>
5    <items n="252" imsqtitype="ims_qtiasiv1p2CC">
6       <qtiitem cod="EUSK1">…</qtiitem>
7       <qtiitem cod="EUSK2">…</qtiitem>
8                              …
9       <qtiitem cod="EUSK252">…</qtiitem>
10   </items>
11   <type>experts</type>
12   <numparameters>1</numparameters>
13   <detailsform>
14     <anchor r="no" />
15     <subtest n="6" nipq="42">
16       <subtest cods="1" nitems="41">
17         <cod>EUSK1</cod>
18         <cod>EUSK7</cod>
19                    …
20         <cod>EUSK241</cod>
21       </subtest>
22       <subtest cods="2" nitems="42"> …</subtest>
23                    …
24       <subtest cods="6" nitems="42"> …</subtest>
25     </subtest>
26     <admtypedata admtype="CALLIE-MOODLE" codexp="expertus" passexp="Exp.123456" />
27     <levelnum>10</levelnum>
28     <nresp>7</nresp>
29     <nexperts>42</nexperts>
30     <correctanswer>yes</correctanswer>
31     <blankresponses>no</blankresponses>
32     <blanklevel>no</blanklevel>
33     <comments>yes</comments>
34   </detailsform>
35   <filtering>
36       <deleteitems>yes</deleteitems>
37       <percentcorrect>70</percentcorrect>
38       <numlevelsrange>3</numlevelsrange>
39       <percentrequest>85</percentrequest>
40       <deleteadmin>yes</deleteadmin>
41       <deleteadminIncorr>yes</deleteadminIncorr>
42       <incomplete>accept</incomplete>
43       <procedure>M.dif</procedure>
44   </filtering>
45 </request>
```

Figura 55 – XML de petición de una calibración tras la especificación del experimento.

El modelo XML generado siempre consta de un elemento raíz petición (*request*). Como *datos generales* (líneas de 1 a 12 en el ejemplo anterior) aparecen el código único de la petición (*code*), el email del responsable que solicita la calibración (*user*), el tipo de calibración seleccionado (*type*), esto es, psicométrica o mediante expertos, y por último el número de parámetros a estimar (*numparameters*). Esta parte contiene también el banco de ítems a calibrar (*ítems*), del que se indica su número de ítems (*n*) y el formato IMS QTI en el que fueron introducidos al sistema (*imsqtitype*). Para cada ítem (*qtiitem*) se incluye un código del propio sistema (*cod*) y el ítem en el formato indicado, que puede ser cualquiera de las versiones del estándar IMS QTI.

El *código único* de la calibración permite identificar cada calibración en curso dentro del sistema y está formado por la concatenación de los siguientes elementos: *REQ* de request, *E* ó *P* dependiendo de que el tipo de calibración solicitado sea mediante expertos o psicométrica, el nombre del *tema* correspondiente al banco de ítems sobre el que se realiza la petición y *la marca temporal* – con la fecha en formato *yyyymmdd* y la hora en formato *hhmmss* – a la que el sistema aceptó dicha petición. Por ejemplo, REQE_EUSK12_20150210_091706 significa que es una petición de calibración mediante expertos sobre el banco EUSK12 aceptada por el sistema el 10 de febrero de 2015 a las 9:17:06 horas. Mediante este formato se garantiza que no haya dos peticiones con el mismo código, ya que el sistema implementado no puede aceptar dos peticiones del mismo tipo y sobre el mismo banco simultáneamente.

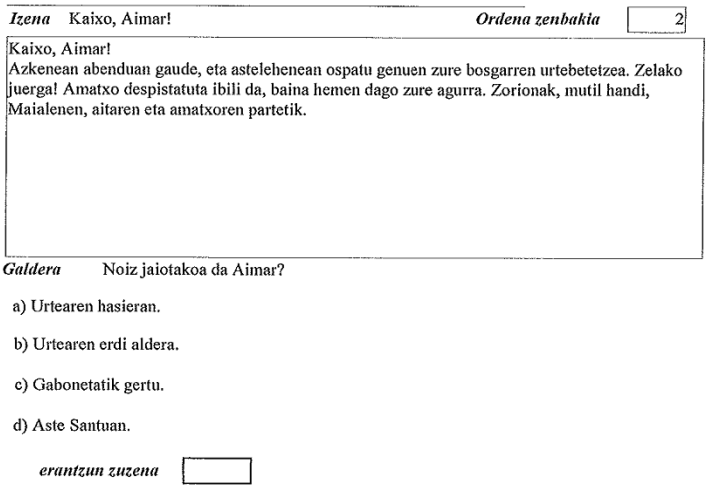
En la parte de detalles para los formularios (*detailsform* líneas de 13 a 34 en el ejemplo) se describen los datos para la administración de los cuestionarios a los expertos, esto es, sobre el diseño de los formularios a administrar y sobre las cuentas de los propios expertos. También se indica el subsistema de administración de cuestionarios que se utilizará (*admtype* de *admtypedata*).

Por último, en la parte filtrado (*filtering* líneas de 35 a 44 en el ejemplo) del modelo XML se indicarán las decisiones tomadas para el tratamiento de la muestra recogida después de la administración de los distintos cuestionarios, esto es, las decisiones para el filtrado de administraciones e ítems durante la fase de análisis y los cálculos de calibración.

A2 Experimento 1: Calibración de los ítems del tema IRALE

La idea de este experimento surge en el momento en que un grupo de miembros de IRALE en Donostia, que trabaja en torno a temas de autoevaluación, contacta con el grupo GHyM por su trabajo en el desarrollo de TAI para el sistema Hezinet. Tras varias reuniones entre los miembros de IRALE y los miembros de la UPV/EHU Armendariz y López-Cuadrado, que trabajan en CALLIE, se diseña un experimento en el que el objetivo final es calibrar psicométricamente un banco de ítems con el cual poder evaluar a los alumnos sobre la *comprensión lectora del euskera* siguiendo el método indicado en (López-Cuadrado et al., 2010). Este método precisa disponer previamente de estimaciones para los valores de la dificultad de cada ítem, lo que se consigue *realizando previamente otra calibración en la que se utilice el juicio de expertos*. Para poder llevar a cabo esta calibración previa, debía estar preparado el banco de ítems a calibrar y haberse definido el modo de evaluar la comprensión lectora.

La preparación y creación del banco de ítems fue realizada por el grupo de trabajo de IRALE junto con algunos otros profesores de IRALE y un experto en materiales de aprendizaje. El banco final contuvo *132 ítems* de selección múltiple. En cuanto a la composición del banco, la mayoría de los ítems (121) contenía una pregunta directa sobre un texto, cuya respuesta correcta había que seleccionar, otros (4) mostraban el texto con un hueco, representado mediante un conjunto de puntos que no estaba situado necesariamente al final, de manera que solo una de las opciones de respuesta se ajustaba adecuadamente, en algunos otros (3) las opciones eran imágenes y en el cuarto tipo de ítem la pregunta estaba formulada en términos negativos (4). Como se muestra en la Figura 56, cada ítem a calibrar consta de una cabecera con el texto a ser leído y un cuerpo con una pregunta de selección múltiple con cuatro opciones de las que solo una es correcta.


Izena Kaixo, Aimar! *Ordena zenbakia* 2

Kaixo, Aimar!
Azkenean abenduan gaude, eta astelehenean ospatu genuen zure bosgarren urtebetetzea. Zelako juerga! Amatxo despistatuta ibili da, baina hemen dago zure agurra. Zoríonak, mutil handi,
Maialenen, aitaren eta amatxoren partetik.

Galdera Noiz jaiotakoa da Aimar?

a) Urtearen hasieran.

b) Urtearen erdi aldera.

c) Gabonetatik gertu.

d) Aste Santuan.

erantzun zuzena

Figura 56 – Formato de un ítem de IRALE para evaluar la capacidad "comprensión lectora".

Por su parte, para obtener una clasificación de los ítems, el grupo de trabajo de IRALE utilizó las escalas descriptivas generales proporcionadas por el Marco Común Europeo de Referencia para las lenguas o CEFR (Council_of_Europe, 2001) para la lectura de un idioma. Concretamente, la autoevaluación de la comprensión lectora está directamente relacionada con el estilo y la complejidad del texto a ser leído, de modo que uno de los parámetros para clasificar el banco de ítems fue el tipo de texto escrito en la cabecera del ítem y el otro fue el nivel de dificultad. En cuanto al *tipo de texto* a ser leído, se utilizaron 7 variantes: narrativo (*NAR*-narrazioak), explicativo (*AZAL*-azalpenak), argumentativo (*ARG*-argudio testuak), descriptivo (*DES*-deskripzioak), instruccional/procedural (*INS*-instrukzio testuak), literario (ANTZ-testu antzerki-bat) y funcional (FUN-funtzionalak). En este último grupo se incluyeron los ítems con imágenes, notas, cartas, diálogos cortos, etc. Por otro lado, para medir la *dificultad de los ítems* se utilizó la escala de 6 niveles comunes recomendada por el CEFR. Dicha escala parte de una división inicial en tres niveles amplios para el usuario – A básico (Basic User), B independiente (Independent user) y C competente (Proficient user) – de la que se realiza una división más fina en A1 (acceso), A2 (plataforma), B1 (umbral), B2 (avanzado), C1 (dominio operativo eficaz) y C2 (maestría). En el caso de IRALE, se añadió un nivel más de refinamiento subdividiendo de nuevo los niveles en una *escala de 11 niveles coherentes*, más o menos del mismo tamaño, es decir, la escala tenía pasos entre A1 y A2, A2 y B1, B1 y B2, B2 y C1, y entre C1 y C2, concretamente aquéllos nominados con el símbolo + como se muestra en la Figura 57.

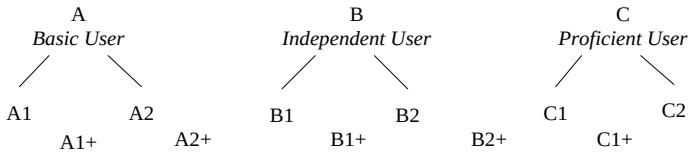


Figura 57 – Escala CEFR de 11 niveles utilizada para la calibración mediante expertos.

Previamente a la calibración, el grupo también hizo una clasificación inicial de los 132 ítems por tipo de texto, obteniendo 1 ANTZ, 21 ARG, 31 AZAL, 7 DES, 25 FUN, 6 INS, 30 NAR y 11 sin clasificar por falta de consenso. A continuación, clasificaron los ítems *grosso modo* por nivel de dificultad obteniendo 47 ítems a nivel *básico* (de A1 a B1), 44 a nivel *avanzado* (B2) y 41 a nivel *competente* (de C1 a C2). La siguiente tabla muestra la clasificación de los 132 ítems del banco. Contiene los ítems ordenados por dificultad y clasificados por tipo de texto, cada uno de ellos con su número identificativo (*Item#*), título (*Título*) y el tipo de texto asociado (*Tipo texto*).

Item#	Título	Tipo texto	Item#	Título	Tipo texto
1	Kaixo, lagun!	FUN	**67**	Eguraldia	FUN
2	Kaixo, Aimar!	FUN	**68**	Ibili eta ibili	NAR
3	Txokolatearen jatorria	AZAL	**69**	Kilo bat tomate	0
4	Txokolatearen museoa	FUN	**70**	Beroari aurre egiteko	AZAL
5	Kaixo, Ane	FUN	**71**	Egunero etortzen zen etxera	NAR
6	Oreina	AZAL	**72**	Hauetako bat EZ da lan-eskaria. Zein?	FUN
7	Haur txokoa	FUN	**73**	Joan den mendean	AZAL
8	Egun on, zer jarriko dizut?	FUN	**74**	Agur t´erdi ikasketa-bidaia	ARG
9	Beste baloiak bezalakoa	DES	**75**	Surflaria ligoi ospea	ARG
10	Kobenkoba	FUN	**76**	BOTEILOIA DEBEKATU	ARG

Item#	Título	Tipo texto	Item#	Título	Tipo texto
11	Agur, andereño	FUN	**77**	Ikerketa asko egin dituzte	AZAL
12	Erosoago egon	0	**78**	Hego Euskal Herrian	ARG
13	Peru nire lagun handia da	DES	**79**	Geltokiko aulkian	NAR
14	Donostiako hipodromoa	FUN	**80**	Gertaerak eta txisteak	NAR
15	Euria eta haize zakarra	AZAL	**81**	Oraindik ere Gipuzkoako herrixka	AZAL
16	Zein EZ da autoan ibiltzeko aholkua?	INS	**82**	Argazkiak	0
17	Kakaoak bide luzea	AZAL	**83**	Infojuego	AZAL
18	Kaixo, Alaitz	FUN	**84**	CAF-Elhuyar sariak	AZAL
19	Aspaldi-aspaldian	NAR	**85**	Kutixiak donostiar erara	DES
20	Donostiako Aquariuma	FUN	**86**	Gödel	AZAL
21	Udazkenean haize handia	NAR	**87**	Herensugea	DES
22	Horoskopoa		**88**	Emakumeak zinema munduan	ARG
23	Ederra, e?	0	**89**	Josten dendan	NAR
24	Artzaina eta Martxoa	NAR	**90**	Demokrazia Churchill	ARG
25	Bidean galtzeko aitzakiarik gabe2	AZAL	**91**	Koartada	0
26	Bidean galtzeko aitzakiarik gabe 1	AZAL	**92**	Urrearen balioa duten hiru puntu	ARG
27	Furgonetan sartu	NAR	**93**	Durrell, biologo eta idazlea	AZAL
28	Berez, Joxe Manueli zegokion	DES	**94**	Muniaren argazkia	
29	DENDA TXIKIAK	ARG	**95**	Gaur egun handi bat izan da	
30	Kaixo, lagunak.	FUN	**96**	CORIUM	AZAL
31	Deabruak eramana	NAR	**97**	Irakasle berria	
32	Guztiz konbentzitu nauzu, Nagore!	ARG	**98**	Humanismoa	ARG
33	Zure gazte txartela eskuratzeko	INS	**99**	BARDOTEN ANTZA	NAR
34	Buruz ikasteko orduan	INS	**100**	Nora, Elena eta mutil bat	ANTZ
35	GRAFITIAK	AZAL	**101**	Nahigabetzen nauenak	ARG
36	Glamourra	ARG	**102**	Amaren esnearekin	AZAL
37	Gizon sendoa zen	NAR	**103**	Bilbao-New York-Bilbao	AZAL
38	Ingalaterran emaniko hiru urteen ostean	NAR	**104**	Edertasuna	0
39	Ia denaren historia labur bat	AZAL	**105**	AITORPENA	0
40	Guztiz ados zurekin, Maialen	ARG	**106**	Artikutza	FUN
41	INGURUKO POBREZIA	ARG	**107**	Zinema	AZAL
42	2001ean, IgNovel sari bat	AZAL	**108**	Irudiaren boterea	AZAL
43	Neke kronikoa	AZAL	**109**	Deskontu.com	AZAL
44	Koala	ARG	**110**	Sendabelarrak	0
45	Haurra “erabat” integratzeko	DES	**111**	e-Mintza	AZAL
46	POLTSA KAKA	ARG	**112**	Mitxelenarena	ARG
47	ANIMAZIOZKO HARRIBITXI ILUNA	AZAL	**113**	Arreba txiki bat dut, eta zer?	ARG
48	LEGATZ FRESKOA	NAR	**114**	Ahizpa biki txalapartariak	NAR
49	ESTEBAN WERFELL	DES	**115**	Gaur igandea da	NAR
50	Kafe Antzokia	ARG	**116**	Zerua marrazteko	NAR
51	35 KILO ESPERANTZA	NAR	**117**	Kontatu didate	NAR
52	NAHI ETA NAHIEZ	0	**118**	Itziar Lazkano	DES

Item#	Título	Tipo texto	Item#	Título	Tipo texto
53	Oparotasunean ohitu gara	ARG	**119**	Pasahitz seguruagoa	NAR
54	MONA LISA	AZAL	**120**	Zorionak, Unax!	INS
55	Proiektuaren helburua	AZAL	**121**	Edurrek amets	FUN
56	Bai. Betiko dira ezizenak Belainen	NAR	**122**	Bi sator txiki	NAR
57	Bidane	NAR	**123**	Ehiztariak eta fruitu-biltzaileak	NAR
58	Tourrean	ARG	**124**	Gero eta tristeago dago	AZAL
59	Neska normala eta zentzuduna	0	**125**	Ipotx urdina	NAR
60	BORDAXAR BASERRIA	NAR	**126**	Saguzarrak	INS
61	Hauetako jarduera bat EZ	FUN	**127**	Txerritxo	AZAL
62	Udalekuetako izen-ematea	FUN	**128**	Ken Zazpi	NAR
63	Kaixo, Maddi	FUN	**129**	Bainujantzia nola zaindu	ARG
64	Andolin Irakin	NAR	**130**	Jakingo duzuenez	INS
65	Talde berezientzako "Lurraldebus" txartela	FUN	**131**	Igande goizean	FUN
66	Ieepa, hobbit maiteok!	FUN	**132**	Larunbatean argazki kamera	FUN

Tabla 25 – Ítems del banco de IRALE clasificados por los responsables.

Para llevar a cabo la calibración manual vía expertos de su banco de ítems, se formó un grupo de miembros de IRALE con 8 integrantes: dos de ellos son los expertos principales y serían también los responsables del proceso de calibración, y los otros 6 fueron reclutados por los anteriores para obtener una calibración vía expertos del banco de ítems. Atendiendo al grado de especialización de este último grupo había dos perfiles de expertos: los que eran especialistas en los niveles más altos de la escala (4) y los que lo eran en los niveles más bajos (2). Además todos ellos trabajaban como profesores de euskera con conocimientos del CEFR y seis de ellos tenían una experiencia superior a los 25 años como profesores o realizando tareas de evaluación en euskera.

Los dos responsables de la calibración manual de IRALE decidieron que cada uno de los dos expertos principales cumplimentaría un primer tipo de cuestionario (*Todos*) con todos los ítems del banco. También diseñaron dos tipos de cuestionarios más, esta vez basados en la clasificación previa de los ítems y el perfil de los 6 expertos restantes: uno de nivel bajo con 85 ítems (*Bajo*) y dificultades comprendidas entre A1 y B2, y otro de nivel alto con ítems de nivel B2 o superior. En los dos tipos de cuestionarios había 26 ítems comunes de nivel avanzado (B2). En el diseño del cuestionario de nivel alto se definieron dos variantes: una con 85 ítems (*Alto.85*) y otra con 67 ítems (*Alto.67*). Estos tres tipos de cuestionarios se asignaron a los 6 expertos participantes teniendo en cuenta su nivel de especialización, es decir, a 2 expertos el Bajo, a 2 expertos el Alto.85 y a los otros 2 expertos el Alto.67.

Mediante email se envió una carta a los expertos antes de empezar a administrar los cuestionarios, y después se les remitió la colección de ítems que debían cumplimentar, con un plazo de 3 semanas para contestar. Durante este tiempo de administración, los expertos responsables siguieron el progreso del trabajo y aclararon las dudas por vía telefónica. Es necesario subrayar que no se estableció ningún criterio común previo ni instrucciones sobre cómo realizar el trabajo de calibración y que cada experto realizó la actividad asignada en solitario.

Respecto a los 8 sujetos que participaron en la calibración, no hubo abandonos. La mayoría de expertos no intentaron todo ni acertaron todas las respuestas correctas,

aunque todos menos uno estuvieron dentro del rango mínimo de respuestas acertadas. En la Tabla 26 se puede ver un resumen de las contribuciones por experto proporcionada por los responsables de IRALE, que además son los dos primeros expertos que aparecen en la tabla.

	experto1	**experto2**	**experto3**	**experto4**	**experto5**	**experto6**	**experto7**	**experto8**
Tipo de cuestionario administrado	Todos	Todos	Bajo	Alto.85	Alto.67	Alto.67	Alto.85	Bajo
Total a calibrar	132	132	85	85	67	67	85	85
Items intentados	132	132	82	80	67	62	85	74
Juicios eliminados	0	0	7	3	0	3	13	21
Juicios aceptados	132	132	75	77	67	57	71	56
%Aceptado	100,00	100,00	88,24	90,59	100,00	85,07	83,53	65,88

Tabla 26 – Resumen de contribuciones por experto.

Además, los responsables de IRALE facilitaron un informe una vez recogidos los datos. En la Tabla 27 se detalla la muestra de datos recogida y revisada por los responsables del experimento. En la primera columna de esta tabla se indica el número identificativo (*Item#*) de los 132 ítems de IRALE. En las 8 columnas siguientes (e*xpertoX*) se recoge el nivel CEFR asignado por el experto X a ese ítem durante la conducción del experimento junto con un * para marcar las respuestas incorrectas.

Item#	**experto1**	**experto2**	**experto3**	**experto4**	**experto5**	**experto6**	**experto7**	**experto8**
1	A1+	A1	A2					A1
2	A1+	A1	A2					A2
3	A2	A2	A2+					A2
4	A2	A1	A1					A1
5	A1+	A2	A1+					A1
6	B1+	B1	B2					B1 *
7	A1+	A1	A2					A1
8	A1+	A1	A1					A1
9	B1+	B1	B1					B1
10	B1+	A2	B1					A2
11	A2	A2+						
12	B1+	B1						
13	A2	A2	A1+					A2
14	B1+	B1	B1					A2
15	A2	A1+						
16	B1	B1+	A2	B2			B2	B1
17	B1	B1	B2					A2
18	A2	A2	A2					A2
19	A2+	A2	B1 *					B1
20	B1+	B1	B1+					B1
21	A2+	B1	A2					B1
22	B1	B1	B1					B1
23	B1+	B1+	B1+	B2			B2	B1
24	B2	B1+	B1	C1			B2	
25	B2	B2	B2	B2	B2	B2	B2	B1
26	B1+	B1+	B1	C1			C1*	A2
27	B1+	B1+	B1+	C1			B2	A2

Item#	experto1	experto2	experto3	experto4	experto5	experto6	experto7	experto8
28	B2	B2	B2	B2+	B2	B2	B2	B2
29	B2	B2	B1	B2	B2	B2	B2 *	
30	B2	B2	B2	B2	B1	B2 B2+	B2	B1
31	B2+	B2+		C1	B2	B2	C1 *	
32	B2	B2+		B2	B2	C1	C1	
33	B2	B2	B1 *	B1+	B1+	B2	B2	A2
34	B2	B2	B1	B2 *	B1+	B2	C1	B1
35	B2+	B2	B2	B2+	B2	B2	B2 *	A2
36	B2+	B2+		B2+	B2+	C1	B1+	
37	C1	B2+		C1	B2	B2+ C1	B2	
38	B2	B2+		C1	B2+	C1	B1+ *	
39	B2+	B2	B1+	B2+	B2	B2+ C1	B1	B1
40	B2	B2	B1 *	B2+ *	B2 B1+	B1+ B2	B2	B1
41	B2+	C1		B2	B2	B1+ B2	B1+	
42	B2+	B2+		B2	B2	C1	B1+	
43	C1	B2+		C1	B2	C1	C1	
44	C1	C1		C1 *	B2	C1	B1+	
45	C1	B2+		C1	C1	C1 C2	B1+	
46	C1+	C1		C1	C1	C1	B2	
47	B2+	B2+		C1	C1		C1	
48	B2+	B2	B1	C1	C1	B2	B2	B1
49	B2+	B2+		C1	B2	C1	B2	
50	C1	B2+		B2+	B2+	C1	B2	
51	B2	B2	B1	B2	B2	B2	B2	B1
52	B2+	B2	B2	C1	B2	B1 B2	B1+	B1
53	B2+	B2	B2	B2 *	B2+ *	C1 *	B2 *	B2
54	C1	C1		C1	B2+	C2	B2	
55	C1	B2+		C1 *	B2+	C2	C1	
56	B2+	B2+		B2	B2	C1	B1+	
57	B2+	C1		B2	B2+	B2	B2	
58	C1	C1		B2+	B2+	C1	C1	
59	B2+	B2	B1	C1	B2+	C1	B2	B1
60	C1	C1		C1	B2+	C1	B2	
61	B1	B1	A2					A1
62	B1+	B1	A2					A1
63	A2	A2	A2					A1
64	B1	A2	B1 *					B1*
65	A2+	A2+						
66	A2+	B1	A2					A1
67	A2+	A2+						
68	B1	B1	A2					A2
69	A1+	A1	A2					A1
70	B1	B1+	B1	B2			B1+	
71	B1+	B1+	A2	B2+			B2	B1
72	B1	B1+	A2	B1			B1+	A1
73	B2	B2+		B2	B2	C1 *	C1 *	
74	B2	B2	B2	B2	B2 *	B2 *	B2	A2+*
75	B2	B2	B1 *	B2	B2+		B1+	
76	B2	B2+		B2+	B2	B2	B2	
77	C1	C1		B2	B2+ *	C1	B2	

Item#	experto1	experto2	experto3	experto4	experto5	experto6	experto7	experto8
78	C1	C1		B2+	B2+	*	C1 *	
79	C1	B2+		C1 *	B2+ C1	C1	B2 *	
80	C1	C1		C1	B2+ *	B2	B2 *	
81	C1	C1		B2+ *	B2	C2 *	B1+	
82	B2	B2		B2+ *	B2		B2	
83	B2	B1+	B2	B1			B2	A2
84	B2	B2	B2	B1+	B2 B1	B2 C1	B2	A2
85	B2	B2+		B2 *	B2+	B2 C1	C1	
86	B2+	B2+		B2+	C1	B2	B2	
87	B1+	B1+	B2				B2	B1
88	B2	B2	B1	B2	B2	B2 C1	B1+	B1
89	B2	B2	B1	B2	B2+	B2 C1	B1+	B1
90	C1	C1		C1	C1		B1+	
91	C1	B2	B1	C1 *	C1	C1 C2	B2	
92	C1	C1		B2+ *	C1	C2	B2	
93	C1	B2+		C1	C1	C1 C2	B1+	
94	A2+	B1	A2					A2
95	A2+	B1	A1+					A1
96	C1	C1		B2	B2+	C2	B2	
97	B1	B1	A2					A1
98	C1	C1		C1 *	B2+	C1 C2	B2	
99	B2	B2	B1	B1	B2 C1		B1+	A2
100	B2	B2	B2	B2+	B1 B2	C1	B1+	
101	C1	C1		C1	C1	C2	C1	
102	B2+	B2	A2	B2	B2+ B2	C1	B2	A2
103	B2+	C1		C1	B2+	C1	C1	
104	C1	C1+						
105	B2+	C1		B2	B2+ C1	C1	B1+	
106	B2+	B2+		B2	B2 B2+	B2	B2	
107	B2	B1+	B1	B2			B1+	A2
108	C1	C1		B2+	B2+	C1	B2	
109	B2	B2	B2	B1+	B2+	B2	B1+	B1
110	C1	B2	B1	B2	C1	*	C1 *	A2 *
111	C1	B2	B1+	C1	C1	C1 C2*	B2	B1 *
112	C1+	C1		C1+	C1 *	C2 *	C1 *	
113	B1+	B1	A2					A2
114	B2	B1+	B1	B2			B1+	B1
115	B1+	B1	A2					A2
116	A2	B1	A1					A1
117	B1	B1	A2					B1
118	B1	B1	B1					A2
119	B1	B1	A2					A2
120	A1+	A1	A2				B1+	A2
121	A2+	B1	A2 *					A2
122	A2+	B1	A1+					A2
123	B1	B1						
124	B1	B1	B1					A2
125	B1	B1						
126	B1	B1+	A2	B2			B2	A2
127	B1	B1	A2					A2

Item#	experto1	experto2	experto3	experto4	experto5	experto6	experto7	experto8
128	B1+	B1+	B1	B2			B1+	A2
129	B1	B1						A2
130	A2+	B1	A2				B1+	A2
131	A2	A2	A1+				B2	A2
132	A1+	A2	A1+				B2	A2

Tabla 27 – Informe de los datos recogidos por los responsables de IRALE.

Como los datos recogidos respecto al nivel de dificultad (ver Tabla 27) fueron muy dispersos y los responsables previeron que sería casi imposible alcanzar un consenso sobre el nivel del ítem hicieron una revisión. Así que una vez recogidos los datos, *se revisó cada contribución de un experto dos veces* y los dos expertos principales decidieron si estos juicios eran aceptados, marcados o eliminados. Durante la primera revisión, *todo juicio de un ítem con respuesta en blanco así como aquéllos con la dificultad omitida o con valores extremos fueron marcadas* en la muestra inicial. Solamente *se eliminó una aportación cuando la discrepancia era de un solo experto y la resolución que dio estaba lejos de la de los otros expertos*. Asímismo, existían 17 valoraciones en las que el experto eligió varios niveles y que los expertos principales convirtieron en un único nivel. Tras este primer filtro, *en una segunda revisión se estudió la respuesta correcta* dada a cada ítem por los expertos y el responsable/creador del ítem decidió retirar el ítem número 125.

Para finalizar, cada nivel del CEFR se convirtió a un valor numérico entero entre 1 y 11, se realizó una media ponderada para cada ítem y se obtuvieron las calibraciones en dificultad para los 131 ítems que permanecían en el banco. Después de obtener los primeros resultados, los responsables dieron un último paso: clasificar los ítems de forma manual según la escala del CEFR. A la vista de los datos recogidos de los expertos y de los resultados obtenidos, los responsables de IRALE consideraron que existían 15 ítems que podían pertenecer a dos niveles distintos: a A1+ o A2 el ítem 132, a A2 o A2+ los ítems 61 y 62, a A2+ o B1 los ítems 64, 66, 94, 97, 121 y 122, a B1+ o B2 los ítems 24, 33 y 34 y a B2+ o C1 los ítems 55, 80 y 111. Para eliminar estos casos de ambigüedad, los responsables analizaron los resultados obtenidos y finalmente se asignó uno de los 11 niveles del CEFR a cada uno de los 131 ítems que permanecieron en el banco, *de modo que hubiese el mismo porcentaje de ítems en los tipos A y C*. En la Tabla 28 se detalla la distribución final de los ítems del banco y como se aprecia en ella, el 16% de los ítems (21 ítems) se clasificaron en el nivel A, el 68% (89 ítems) en el nivel B y el 16% (21 ítems) en el nivel C.

Niveles	A1	A1+	A2	A2+	B1	B1+	B2	B2+	C1	C1+	C2
Nº de ítems	7	1	10	3	29	14	26	20	20	1	0

Tabla 28 – Distribución final de los ítems del banco según los 11 niveles del CEFR.

Tras la calibración manual realizada por los 8 miembros citados del grupo IRALE sobre su banco de 132 ítems, se ideó el experimento 1, con objeto de calibrar un banco de ítems de idénticas características y proceso de calibración mediante CALLIE-EXPERT y que se aprovechó para iniciar la evaluación de los principales componentes de CALLIE. Este experimento se prolongó desde febrero hasta junio de 2013. En el

resto de este anexo, se presentan los detalles del mismo, desarrollado por la autora de esta memoria.

En este experimento participó un grupo de 8 profesores de la UPV/EHU, que son usuarios potenciales de CALLIE, conocedores de temas de evaluación y calibración pero no de herramientas informáticas semejantes a CALLIE. Todos conocen la plataforma Moodle, pero no en profundidad en lo relativo a la creación de grupos de alumnos ni a la administración de cuestionarios y siempre actúan con ella con el rol de profesor. Además, los dos profesores que se captaron como responsables eran filólogos vascos con conocimientos básicos sobre el CEFR y uno de ellos tenía una experiencia superior a los 20 años como profesor en euskera. La idea era que estas 8 personas adoptaran los mismos roles que cada uno de los integrantes del grupo que realizó la calibración manual.

Como recursos para que los participantes pudieran hacer este experimento, se instaló en modo local – utilizando un portátil de la UPV/EHU – la aplicación CALLIE-EXPERT, que se puso a disposición de los dos profesores que actuarían como responsables. El componente CALLIE-MOODLE fue alojado en el servidor Windows del departamento de Lenguajes y Sistemas Informáticos y estuvo disponible para todos los participantes.

El experimento se compone de cuatro pruebas sucesivas: prueba 1 (Creación del banco de ítems con ESKARI), prueba 2 (Diseño del experimento con ESKARI), prueba 3 (Administración de ítems con CALLIE-MOODLE) y prueba 4 (Recogida de datos, análisis y calibración con PRO-C). Además, las cuatro pruebas aplicadas sucesivamente llevan a una calibración del banco de ítems que posteriormente se podrá replicar.

Ante la imposibilidad de captar sujetos con una experiencia tan dilatada como los del grupo IRALE, a los dos profesores que actuarían como expertos principales se les ofrecería ayuda en temas de CEFR y calibración, y a los otros 6 profesores, que actuarían como expertos reclutados siguiendo el papel de los 6 expertos originales, se les facilitarían las respuestas que dieron aquéllos. Asímismo, por motivos de confidencialidad, la desarrolladora solamente tendría acceso a los datos presentados en las cuatro tablas anteriores, por tanto, no conocería los datos concretos de los ítems (ni texto a leer ni opciones) ni a las administraciones del grupo de IRALE. Por todo ello, antes de dar comienzo al experimento, la desarrolladora del experimento adaptó y completó las decisiones de diseño para que concordasen con las que se habían realizado manualmente por parte de IRALE y con los datos y recursos humanos disponibles.

Concretamente, para la prueba 1 se decidió preparar un banco de ítems para el experimento llamado IRALE que contuviera 132 ítems con las mismas características que los originales (títulos, nivel grosso modo y tipo de texto de la Tabla 25), con identificativo IRALEX siendo X el número identificativo del ítem, y un texto a leer ficticio. En todos los ítems, la primera opción de las cuatro presentadas sería siempre la respuesta correcta.

En cuanto al diseño completo de la calibración para el banco de ítems, necesario para la prueba 2, por un lado, los 4 tipos de cuestionarios elaborados por IRALE se dividieron – igual que en el original – en 5 bloques disjuntos de 47, 26, 12, 6 y 41 ítems respectivamente, de modo que el cuestionario *Todos* de 132 ítems tendrá los 5 bloques, el cuestionario *Bajo* de 85 ítems estará formado por los bloques de 47, 26 y 12 ítems, el cuestionario *Alto.85* constará de los bloques de 26, 12, 6 y 41 ítems y el cuestionario *Alto.67* tendrá los bloques de 26 y 41 ítems. Todos estos datos se resumen en la Tabla 29.

Bloque#	Nº de ítems	Items del banco IRALE	Cuestionarios en los que se encuentra
1	47	de IRALE1 a IRALE47	Todos y Bajo
2	26	de IRALE48 a IRALE73	Todos, Bajo, Alto.85 y Alto.67
3	12	de IRALE74 a IRALE85	Todos, Bajo y Alto.85
4	6	de IRALE86 a IRALE91	Todos y Alto.85
5	41	de IRALE92 a IRALE132	Todos, Alto.85 y Alto.67

Tabla 29 – Bloques e ítems asignados a cada bloque.

Por otro lado, se decidió que todo el grupo participaría como experto en la calibración, que se les administrarían los ítems con CALLIE-MOODLE y que los 8 expertos participantes se identificarían con los nombres de usuario "exp1", "exp2", "exp3", "exp4", "exp5", "exp6", "exp7" y "exp8" respectivamente. Como se refleja en la Tabla 30, cada uno de los dos expertos principales ("exp1" y "exp2" que además son los propios responsables) cumplimentaría el primer tipo de cuestionario (*Todos*) con todos los ítems del banco. El resto de tipos se asignaron a cada experto teniendo en cuenta el nivel de especialización del experto original que había participado en la calibración manual, esto es, a "exp3" y "exp8" se les asignaría el de nivel bajo (*Bajo*), a "exp4" y "exp7" el de nivel alto con 85 ítems (*Alto.85*) y a "exp5" y "exp6" el de nivel alto con 67 ítems (*Alto.67*).

Cuestionario#	Nombre	Nº de ítems	Bloques que lo componen	Expertos que lo deben cumplimentar
1	Todos	132	1, 2, 3, 4 y 5	exp1 y exp2
2	Bajo	85	1, 2 y 3	exp3 y exp8
3	Alto.85	85	2, 3, 4 y 5	exp4 y exp7
4	Alto.67	67	2 y 5	exp5 y exp6

Tabla 30 – Cuestionarios, bloques de cada uno y asignación a expertos.

Se pediría para cada ítem tanto el nivel estimado como la respuesta correcta al ítem, pudiendo el experto omitir cualquiera de ellas, y se permitirán comentarios propios. En cuanto al filtrado de los datos *se aplicará la malla más restrictiva recomendada para los ítems* por la propia herramienta CALLIE, esto es, es posible eliminar ítems del banco cuando no superen el 70% de respuestas correctas o bien el número de valoraciones en una horquilla de 3 niveles de los 11 niveles posibles de dificultad considerados sea inferior al 85% del total. Por último, como el número de expertos disponibles es bajo y se considera que todos ellos son expertos fiables, no se permitirá descartar administraciones enteras, es decir, se mantendrán las administraciones incorrectas y se aceptarán las incompletas.

Para la administración de ítems con CALLIE-MOODLE (necesario en la prueba 3) se *creó un curso en CALLIE-MOODLE* tal y como lo generaría la propia herramienta CALLIE-EXPERT con la definición de la prueba 2. Así, se creó automáticamente el curso en Moodle con las mismas decisiones de ese diseño, en este caso los 5 bloques definidos y el test para aportaciones propias. Los dos responsables de la calibración podrían acceder a este curso como profesores y cada experto podría acceder como alumno con el nombre de usuario y contraseña que se configuraron con CALLIE-ESKARI en este caso de "exp1" a "exp8". Se generaron tres campos por ítem, el primero correspondiente al ítem a calibrar, que es de selección múltiple e incluye el texto a leer antes de la pregunta que el experto deberá contestar. En el caso de la

calibración manual, a los expertos originales se les pidió estimar para cada ítem del cuestionario uno o varios niveles adyacentes entre los 11 posibles. Por tanto, en el segundo campo "zailtasuna" (dificultad, en euskera) el experto del experimento 1 estimaría el nivel principal de dificultad entre 1 y 11 y en el tercero "besteak" (otros, en euskera) podría indicar otros posibles niveles adyacentes. La Figura 58 muestra el mismo ítem que la Figura 56 generado automáticamente por CALLIE-EXPERT en el curso de Moodle para la calibración mediante expertos del banco IRALE, tal y como lo vería uno de los expertos del experimento 1 al entrar en el curso.

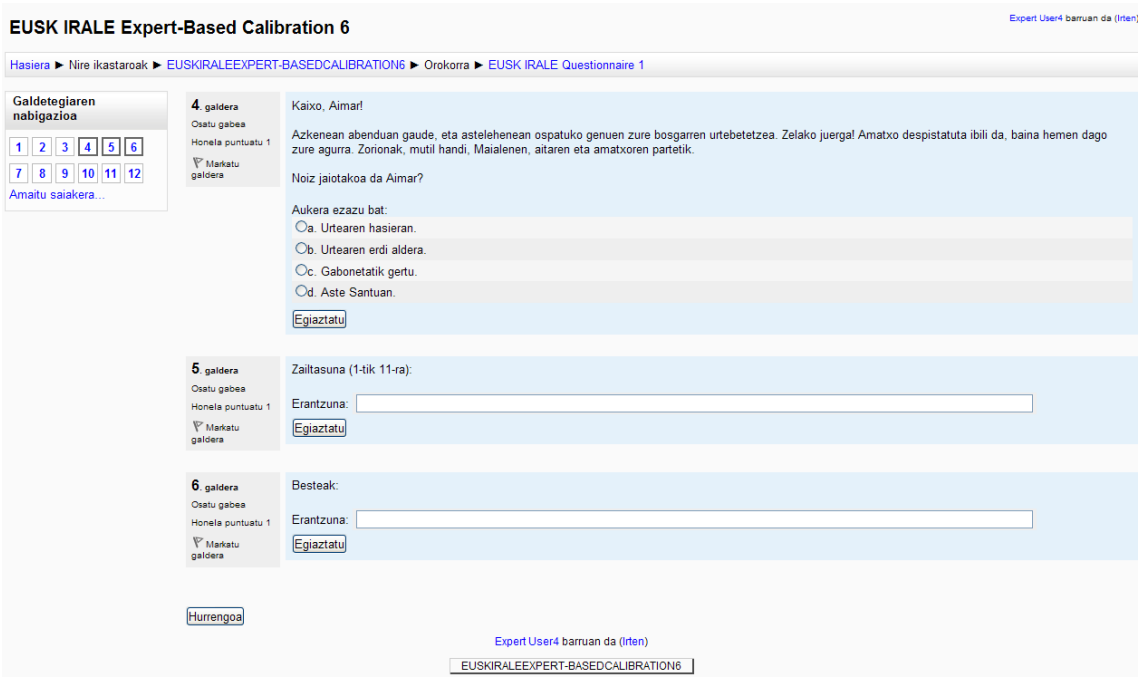


Figura 58 – Datos a rellenar por ítem en la prueba 3 con CALLIE-MOODLE.

En base a las decisiones anteriores, los cuestionarios/bloques a rellenar por experto en este curso se corresponderían con lo indicado en la Tabla 30. Por último, la desarrolladora preparó las respuestas individuales que debía dar cada uno de los 8 expertos participantes, a partir del informe proporcionado por el grupo de IRALE (Tabla 27).

Al experimento se le dedicaron cuatro sesiones: cada una para explicar y ayudar en cada una de las pruebas. En cada sesión se siguieron los mismos pasos: (1) explicar la funcionalidad a utilizar en CALLIE para la prueba, (2) permitir al usuario que interactúe con ella, (3) explicar la tarea a realizar con CALLIE y (4) proponerle que al finalizar cada prueba rellene una encuesta en la que puntúe el componente utilizado en una escala de 1 a 10 y comente la impresión que le ha causado. En las sesiones se trató exclusivamente con los dos profesores que actuarían como responsables, que fueron los encargados de pasar el material y explicar a sus expertos la prueba en la que participarían, así como de recopilar la información sobre el desarrollo de esa prueba. Además de estas sesiones, los responsables podían consultar en cualquier momento dudas sobre la herramienta o el proceso de calibración.

En el resto de este anexo, se detalla para cada prueba: en qué consistió, los participantes, el desarrollo de la prueba y la puntuación otorgada a la funcionalidad utilizada junto con los comentarios relevantes. Además, al final del anexo se ofrecen los

resultados útiles conseguidos con cada una de estas pruebas: por un lado, los resultados intermedios necesarios para los siguientes pasos del proceso de calibración del banco de ítems y, por otro lado, los resultados útiles para evaluar el funcionamiento del componente de CALLIE-EXPERT objeto de la prueba.

PRUEBA 1

Creación del banco de ítems con ESKARI

La prueba consistió en introducir en CALLIE, por orden de dificultad, los 132 ítems ficticios del banco en un nuevo tema – llamado IRALE – con los 7 tipos de texto como categorías y los 3 niveles previos de dificultad preasignados por los miembros de IRALE. Como tarea previa, los dos responsables de IRALE habían clasificado ya los ítems del banco IRALE tanto en tipo de texto como en dificultad *grosso modo* y la desarrolladora había dado las pautas para la creación del mismo.

Participantes y desarrollo de la prueba

En esta prueba participaron los dos profesores de la UPV/EHU que actuaban como responsables de la calibración. Respecto a la creación del tema IRALE y la introducción de los 132 ítems, los dos participantes realizaron la tarea de manera conjunta y correctamente mediante el paso 1 del menú de ESKARI. Para la introducción de los ítems siempre utilizaron la utilidad para ítems de selección múltiple y cada uno introdujo 66 ítems.

Al finalizar la prueba todos los participantes rellenaron la encuesta correspondiente. En cuanto a la puntuación dada por ambos responsables fue de 8. Los dos participantes consideran que el componente es adecuado para su función y que funciona correctamente. Como comentario negativo se indicó que no hay posibilidad de introducir varios ítems a la vez.

PRUEBA 2

Diseño del experimento con ESKARI

La prueba consistió en diseñar mediante CALLIE-EXPERT una petición de calibración idéntica a la que correspondería al experimento de calibración manual para el banco de ítems IRALE. Los pasos de la prueba fueron: (1) identificar las distintas características del diseño, (2) introducir el diseño en CALLIE con el menú de cuatro pasos y (3) ver el resumen de la petición, corregir si es necesario y confirmar la petición comprobando que aparece en CALLIE-PRO. Como tarea previa se había realizado ya la prueba 1, es decir, está creado el tema IRALE en CALLIE-EXPERT con sus 132 ítems.

Participantes y desarrollo de la prueba

En esta prueba participaron los dos profesores de la UPV/EHU que actuaban como responsables de la calibración.

Tras proporcionarles documentación con el diseño completo del experimento, estos dos profesores realizaron sucesivamente los cuatro pasos del menú. En los pasos 1 y 2 del menú todos los participantes seleccionaron adecuadamente el tema IRALE y la calibración mediante expertos. En el paso 3 todos definieron el tipo de administración mediante *CALLIE-MOODLE* y un reparto inicial de ítems *continuo*. Obviamente, en

este caso la opción en CALLIE para el diseño de cuestionarios fue la de cuestionarios no homogéneos ya que los 4 tipos de cuestionarios diseñados tenían distinta longitud. A continuación, siguiendo con el paso 3, indicaron los parámetros específicos para diseñar estos cuestionarios no homogéneos. Los cuatro tipos de cuestionarios diseñados implicaban la división del banco de ítems en cinco bloques de ítems disjuntos, esto es 5 bloques con los distintos ítems y los 8 expertos disponibles. Durante la introducción de estos datos, CALLIE generó automáticamente dos alertas debidas al bajo número de expertos captados y a la posibilidad de no llegar a las 7 opiniones por ítem, que en ningún caso les impidieron seguir con el proceso de calibración. Acto seguido, indicaron la composición de los distintos bloques y cuestionarios en CALLIE-EXPERT. El sistema también emitió dos avisos (en naranja) referidos a la posible duración excesiva y cantidad de opiniones, utilizando una media orientativa que en ningún caso impide seguir con el proceso de calibración. Para finalizar el paso 3, se debían asignar los distintos ítems a cada uno de estos bloques, con los que se crearían posteriormente los cuatro tipos de cuestionarios, idénticos a los diseñados por IRALE. En el paso 4 todos los participantes volvieron a acertar considerando 11 niveles posibles de dificultad y eligiendo para el resto de datos las opciones descritas en la planificación del diseño. Los dos participantes realizaron todo correctamente puesto que podían preguntar si tenían dudas.

Al finalizar la prueba todos los participantes rellenaron la encuesta correspondiente. En cuanto a la puntuación dada por cada uno de los responsables, ambos puntuaron el menú con un 9. Los dos participantes consideran que el componente es adecuado para su función y que funciona correctamente. Como comentario se valora positivamente los valores automáticos de los parámetros.

PRUEBA 3

Administración de ítems con CALLIE-MOODLE

Esta prueba se dividió en dos partes que corresponden a las tareas de administración de ítems, por parte de los responsables y de los expertos respectivamente.

Los *participantes en la prueba como responsables* de la calibración debían conducir la administración de los ítems ideada por los dos responsables de IRALE y que se asocia a la petición de calibración resultante de la prueba 2. Las tareas a realizar consistieron en supervisar y controlar toda la administración de los ítems a los expertos mediante actividades tales como avisar a los expertos de los cuestionarios que deben rellenar, incluir material de ayuda, establecer plazos, resolver dudas, etc. Para realizar estas tareas, se les proporcionó toda la información extraida de la calibración manual respecto a estos temas, contenida en la Tabla 27 y en la Tabla 30, el plazo de tres semanas, las posibles vías de consulta, etc. Aunque se les instó a utilizar Moodle y se les explicaron someramente las posibilidades que ofrecía, se les dio libertad para usar cualquier otro método de comunicación.

Por su parte, cada *participante en la prueba como experto* debía cumplimentar y enviar en plazo mediante CALLIE-MOODLE los cuestionarios que le indicaran los responsables de la calibración utilizando las respuestas predeterminadas que también les habían sido facilitadas.

Participantes y desarrollo de la prueba

En esta prueba participaron los dos profesores de la UPV/EHU que actuaban como responsables de la calibración y todos los componentes del grupo del experimento (8 personas) como expertos.

Antes de comenzar la prueba, la desarrolladora del experimento facilitó a los responsables tanto la asignación cuestionario/bloque/experto (Tabla 31) como las respuestas individuales que debía dar cada experto participante, y les dio las instrucciones oportunas para su difusión.

Experto	Cuestionario#	Nombre Cuestionario	Bloques a cumplimentar
exp1	1	Todos	1, 2, 3, 4 y 5
exp2	1	Todos	1, 2, 3, 4 y 5
exp3	2	Bajo	1, 2 y 3
exp4	3	Alto.85	2, 3, 4 y 5
exp5	4	Alto.67	2 y 5
exp6	4	Alto.67	2 y 5
exp7	3	Alto.85	2, 3, 4 y 5
exp8	2	Bajo	1, 2 y 3

Tabla 31 – Cuestionario a cumplimentar por cada experto del experimento 1.

Los responsables podían proporcionar en CALLIE-MOODLE un documento adicional con todas las instrucciones para una correcta cumplimentación de los cuestionarios, contactar con los demás expertos y monitorizar directamente el progreso de la recogida de datos, así como descartar aportaciones, ítems y expertos. A pesar de ello, no emplearon Moodle ni para indicar instrucciones ni para resolver las consultas de los expertos. En vez de ello, la comunicación se produjo en persona, vía email o teléfono.

En cuanto a los 8 participantes en calidad de expertos, todos ellos rellenaron adecuadamente sus cuestionarios en el curso de CALLIE-MOODLE y en el plazo establecido.

Una vez recogidos los datos en CALLIE-MOODLE, *los dos participantes que actuaban como responsables revisaron cada contribución de un experto aplicando idénticos criterios que en la calibración manual.* Como resultado de esta revisión se marcaron 49 de las 714 contribuciones iniciales de la muestra (ítems intentados). El análisis manual llevado a cabo por los responsables eliminó 47 de estas aportaciones (juicios eliminados) con lo que las aportaciones aceptadas fueron 667 (juicios aceptados). Además, se descartaron otras 2 aportaciones por la retirada del ítem 125. Todo quedó marcado para la recogida de datos mediante CALLIE-EXPERT.

Al finalizar la prueba todos los participantes rellenaron la encuesta correspondiente.

En cuanto a la puntuación dada por cada uno de los responsables, ambos puntuaron CALLIE-MOODLE con un 9. Los dos participantes consideran que el componente es adecuado para su función y que funciona correctamente. Sus comentarios fueron positivos: se juzga que Moodle es una plataforma adecuada para administrar y seguir el progreso de los expertos en la cumplimentación de los cuestionarios.

En cuanto a la puntuación dada por cada uno de los expertos participantes, éstas fueron por orden: 9, 9, 8, 8, 8, 8, 9 y 7. Todos los participantes consideran que el componente es adecuado para su función y que funciona correctamente. Respecto a los comentarios registrados, se juzga que Moodle es una plataforma adecuada para la cumplimentación de los cuestionarios y se refieren problemas al intentar terminar el cuestionario y enviarlo cuando no se ha cumplido el número máximo de intentos.

PRUEBA 4

Recogida de datos, análisis y calibración con PRO-C

La prueba consistió en analizar, calibrar y obtener los resultados detallados de una petición de calibración idéntica a la que correspondería al experimento de calibración para IRALE. Una vez recogidos y revisados los datos cumplimentados por los expertos, el objetivo final de esta prueba sería calibrar los 131 ítems en dificultad utilizando los datos facilitados por los miembros del grupo IRALE, ahora mediante CALLIE-EXPERT. Para conseguir este objetivo, y si se consideraba necesario, se debían obtener resultados alternativos mediante simulación con la herramienta. Los pasos a dar fueron: (1) recoger los datos de las administraciones de ítems del tema IRALE desde CALLIE-MOODLE, (2) analizar y calibrar, (3) ver resultados detallados, y (4) probar y obtener resultados alternativos más adecuados a las necesidades de los responsables en caso de que los iniciales no sean satisfactorios, esto es, si no se consiguen resultados para todos los ítems o se descartan demasiadas aportaciones. Como tarea previa se habían realizado ya las pruebas 2 y 3, con lo que los datos aportados por los expertos estaban almacenados y revisados en el curso correspondiente de CALLIE-MOODLE que se asoció a la petición creada en la prueba 2.

Participantes y desarrollo de la prueba

En esta prueba participaron los dos profesores de la UPV/EHU que actuaban como responsables de la calibración. Estos responsables recogieron los datos desde el curso de CALLIE-MOODLE y los depuraron y calibraron con CALLIE-EXPERT. Finalizado el análisis de los datos, CALLIE-EXPERT halló los 80 niveles de dificultad correspondientes a los ítems no descartados por los análisis y obtuvo un valor real en el intervalo [1-11] para cada uno de ellos. A partir de las hojas de cálculo generadas con los detalles, los responsables estudiaron los resultados obtenidos por CALLIE-EXPERT para cada uno de los ítems del banco, es decir, bien el valor de la dificultad y las aportaciones empleadas, o bien el análisis que lo descartó.

Estos resultados iniciales no fueron aceptables, ni por la cantidad de ítems con la dificultad calculada (80), ni por la cantidad de aportaciones utilizadas (el 57,56%). Así, para conseguir estimaciones de dificultad para todos los ítems del banco los responsables utilizaron la opción de *Simulación* de CALLIE-PRO que permite aplicar distintos análisis a los mismos datos. Tras realizar varias pruebas, se escogió relajar el porcentaje de opiniones hasta el 60% manteniendo la horquilla de 3 niveles y se respetó el resto del experimento original. Con esta nueva malla obtuvieron valores para la dificultad de los 131 ítems no descartados.

Al finalizar la prueba todos los participantes rellenaron la encuesta correspondiente. En cuanto a la puntuación dada por ambos responsables fue de 9. Los dos participantes consideran que el componente es adecuado para su función y que

funciona correctamente. Hubo dos comentarios positivos: los resultados se ven fácilmente y se pueden generar distintos resultados.

RESULTADOS PARA LA CALIBRACIÓN DEL TEMA IRALE

La celebración sucesiva de cada una de las pruebas fue creando una serie de resultados en CALLIE que se describen en los siguientes párrafos.

Como **resultado de la prueba 1** los participantes como responsables generaron el tema IRALE con un banco de ítems en CALLIE para su posterior calibración en el sistema. Los ítems de este tema poseen las mismas características que el banco real preparado por el grupo IRALE en su calibración manual.

Como **resultado de la prueba 2** se diseñó la calibración del tema IRALE y se generó una petición de calibración aceptada por CALLIE-EXPERT para la calibración vía expertos.

Como **resultado de la prueba 3** los datos aportados por los expertos de la calibración manual fueron almacenados y revisados en el curso correspondiente de CALLIE-MOODLE que se asoció a la petición generada en la prueba 2.

Como **resultado de la prueba 4** se obtuvieron con CALLIE-EXPERT las calibraciones en dificultad de los 131 ítems del banco de ítems IRALE como resultado del procesado de la petición.

REVISIÓN DE LAS TAREAS INVOLUCRADAS

Para llevar a cabo la evaluación de cada componente de CALLIE-EXPERT se detectaron las tareas involucradas en cada prueba y se realizó una revisión de las mismas, que se resume en las dos tablas siguientes. La Tabla 32 refleja la participación de los dos responsables de la calibración del banco IRALE y la Tabla 33 refleja la participación de todo el grupo en el rol de expertos durante la calibración de los ítems del banco IRALE. En cada tarea aparece si el participante la realizó correctamente o no (Ok o Mal), salvo en la prueba 3 (parte responsables), en la que los dos responsables completaron las tareas asignadas correctamente, y se revisó la utilización o no de Moodle en la ejecución de las diferentes tareas a realizar durante el desarrollo de la prueba. En la última columna aparece el porcentaje global de realización correcta.

	Tarea involucrada	Responsable1	Responsable2	Grado de realización
Prueba 1	**Creación banco**	Ok	Ok	**100%**
	Introducción ítems	Mediante utilidad (66)	Mediante utilidad (66)	**100%**
Prueba 2	**Pasos 1 y 2**	Ok	Ok	**100%**
	Reparto inicial	Continuo	Continuo	**100%**
	Resto paso 3	Ok	Ok	**100%**
	Paso 4	Ok	Ok	**100%**
	Confirmar Petición	Ok	Ok	**100%**
Prueba 3 (parte responsables)	**Avisos cumplimentación**	Sí	Sí	**100%**
	Administración	Sí (Moodle)	Sí (Moodle)	**100%**

	Revisión datos	Sí (Moodle)	Sí (Moodle)	**100%**
	Instrucciones	No (email, teléfono)	No (email, teléfono))	**0%**
	Resolución de consultas	No (email, teléfono)	No (email, teléfono)	**0%**
Prueba 4	**Recoger datos desde Moodle**	Ok	Ok	**100%**
	Depurar y calibrar	Ok	Ok	**100%**
	Ver resultados detallados	Ok	Ok	**100%**
	Simular	Ok	Ok	**100%**

Tabla 32 – Revisión de las tareas de los responsables en el experimento 1.

	Tarea involucrada	**Exp1**	**Exp2**	**Exp3**	**Exp4**	**Exp5**	**Exp6**	**Exp7**	**Exp8**	**Grado de realización**
Prueba 3 (parte expertos)	**Rellenar cuestionario con Moodle**	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	**100%**
	Enviar cuestionario con Moodle	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	**100%**
	¿Intenta todos sus ítems?	Sí	Sí	No	No	Sí	No	Sí	No	**50%**
	¿Da la respuesta correcta a todos sus ítems intentados?	Sí	Sí	No	No	Sí	No	No	No	**37,5%**
	¿Alcanza el mínimo de respuestas correctas (75%)?	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	**87,5%**

Tabla 33 – Revisión de las tareas de los participantes expertos en el experimento 1.

A3 Experimento 2: Réplicas y evaluación con alumnos

En este anexo se presentan los detalles del experimento 2, desarrollado por la autora de esta memoria, que se llevó a cabo con varios universitarios para la evaluación de los distintos componentes de CALLIE. Este experimento tuvo lugar el 27 de Septiembre de 2013 en los laboratorios de la Escuela Universitaria de Ingeniería Técnica Industrial de Bilbao (UPV/EHU).

En este experimento participaron cinco alumnos de segundo curso del grado de informática. Su perfil es el de usuarios inexpertos en temas de evaluación y calibración, pero conocedores y usuarios de distintas herramientas informáticas.

Para que las pruebas de los participantes no interfiriesen entre sí, cada uno de los ordenadores utilizados dispuso de su propia aplicación CALLIE-EXPERT en modo local, con el componente CALLIE-MOODLE en el servidor Windows del departamento de Lenguajes y Sistemas Informáticos.

El experimento se compuso de cuatro pruebas dependiendo de la funcionalidad y el componente que pretende evaluar: prueba 1 (Creación del banco de ítems con ESKARI), prueba 2 (Diseño del experimento con ESKARI), prueba 3 (Preparación de un libro Excel con datos recogidos para PRO-C) y prueba 4 (Recogida de datos, análisis y calibración con PRO-C). Además, las pruebas 2 y 4 aplicadas sucesivamente replican las calibraciones originales de Hezinet e IRALE y constaron de dos partes, denominadas como la calibración a la que se refieren.

Al experimento se le dedicó una única sesión de tres horas. Una semana antes, se había contactado mediante email con los participantes y se les había proporcionado información sobre el experimento, sobre CALLIE y sus componentes e información básica sobre evaluación y calibración. Respecto al día del experimento, durante la primera hora se hizo una presentación con la explicación del experimento, de CALLIE y de los conceptos que se manejarían en él y se hizo una ronda de preguntas ya que se había establecido no resolver ninguna duda durante el transcurso de las pruebas. Durante las dos horas siguientes los 5 alumnos realizaron las cuatro pruebas, disponiendo de un tiempo máximo de 15, 25, 15 y 30 minutos respectivamente para cada una de ellas. Al finalizar cada prueba el participante rellenaría una encuesta en la que puntuaría el componente utilizado en una escala de 1 a 10 y comentaría la impresión que le ha causado.

En el resto de este anexo, se detalla para cada prueba: en qué consistió, el desarrollo de la prueba y la puntuación otorgada a la funcionalidad utilizada junto con los comentarios relevantes. Además, al final del anexo se ofrecen los resultados útiles conseguidos con cada una de estas pruebas: por un lado, los resultados intermedios necesarios para la réplica de los procesos de calibración de los bancos de ítems de Hezinet e IRALE y, por otro lado, los resultados útiles para evaluar el funcionamiento del componente de CALLIE-EXPERT objeto de la prueba.

PRUEBA 1

Creación del banco de ítems con ESKARI

La prueba consistió en pedir a cada participante que introdujera 10 ítems distintos del banco de ítems de Hezinet en un tema ya existente: PRU_EUSK. Para ello, se les proporcionó para cada ítem a introducir el código, la pregunta (título), las cuatro opciones y la opción correcta de ítems almacenados en la base de datos del experimento original. Los primeros 5 ítems se debían introducir con la página para ítems de selección múltiple (Figura 26 de la memoria) y los otros 5 utilizando la importación en formato IMS QTI apoyándose en la ayuda de la propia herramienta (Figura 25 de la memoria). Como tarea previa la desarrolladora del experimento, había creado el tema PRU_EUSK accesible para todo usuario de CALLIE y preparado varios documentos con los distintos ítems de Hezinet a introducir, que se repartieron entre los participantes.

Para completar esta prueba se estableció un tiempo máximo de 15 minutos.

Desarrollo de la prueba

Respecto a la introducción de los ítems en el tema PRU_EUSK, uno de los alumnos creó un banco nuevo, en vez de utilizar el dado para la prueba y otro no fue capaz de introducir a tiempo los ítems en formato IMS QTI, puesto que no entendió como debía hacerlo a pesar de consultar la ayuda disponible. Todos los demás terminaron correctamente la totalidad de la tarea encomendada.

Al finalizar los 15 minutos de la prueba todos los participantes rellenaron la encuesta correspondiente. En cuanto a la puntuación sobre la introducción de ítems dada por los cinco alumnos en orden fue: 7, 6, 6, 7 y 6. Media: 6,4. Los cinco participantes consideran que el componente es adecuado para su función y que funciona correctamente. Hubo dos comentarios positivos: sobre la introducción de ítems con la utilidad, y la existencia de ayuda sobre el formato IMS QTI para el ítem. No hubo comentarios negativos.

PRUEBA 2

Diseño del experimento con ESKARI

Esta prueba se dividió en dos partes que correspondieron al diseño de las calibraciones para el banco de Hezinet y para el banco de IRALE respectivamente.

La **parte Hezinet** de la prueba consistió en diseñar una petición de calibración idéntica a la que correspondería al experimento de calibración para Hezinet y confirmar la petición comprobando que está en CALLIE-PRO. Las características del diseño a especificar fueron: utilizar el banco de ítems EUSK, para realizar sobre él una calibración mediante expertos según los 12 niveles de HABE, contando con 116 expertos disponibles y con una administración de ítems externa al sistema CALLIE-MOODLE, con todas las demás opciones las recomendadas por defecto por la herramienta. Por último, tomar las mismas decisiones de filtrado y administración que en el experimento original, esto es, utilizar una escala de 12 niveles, pedir obligatoriamente para cada ítem tanto el nivel como la respuesta correcta y permitir comentarios propios. En cuanto al filtrado de los datos permitir borrar ítems cuando no

superen el 50% de respuestas correctas o bien el número de valoraciones en una horquilla de 4 niveles de los 12 posibles sea inferior al 75% del total, permitir borrar administraciones incorrectas y aceptar las incompletas.

Como tarea previa para la parte de Hezinet se creó el banco de ítems a calibrar en el sistema CALLIE. Para ello, se generaron automáticamente en CALLIE los 252 ítems de Hezinet en formato IMS QTI empleando para cada ítem el código, la pregunta (título), las cuatro opciones y la opción correcta almacenados en la base de datos del experimento original. Además, todos ellos se englobaron en un mismo tema denominado EUSK.

La **parte IRALE** de la prueba consistió en diseñar una petición de calibración idéntica a la que correspondería al experimento de calibración para el banco de ítems IRALE como se decidió en el experimento 1 y confirmar la petición comprobando que está en CALLIE-PRO. Las características del diseño a especificar fueron: utilizar el banco de ítems IRALE, para realizar sobre él una calibración mediante expertos según los 11 niveles del CEFR, contando con 8 expertos disponibles a los que se les administrarán los ítems con CALLIE-MOODLE y datos de acceso con los nombres de usuario personalizados como "expNº" donde Nº será el número asignado automáticamente por CALLIE-EXPERT. Respecto al diseño de los distintos cuestionarios a administrar, es el descrito en las tablas del experimento 1. Por último, en cuanto a las decisiones de filtrado y administración, no se permite eliminar administraciones, se pide a cada experto el nivel y la respuesta al ítem pero se le permite dejar cualquiera de ellas en blanco.

Como tarea previa para la parte de IRALE mediante el experimento 1 se había creado el tema IRALE con sus 132 ítems. Además, también se había decidido todo lo referente al diseño del experimento para esa calibración.

Para completar esta prueba se estableció un tiempo máximo de 25 minutos.

Desarrollo de la prueba

En cuanto al desarrollo de la parte Hezinet, los cinco alumnos realizaron sucesivamente los cuatro pasos del menú. En el paso 1 del menú 3 alumnos seleccionaron los ítems del tema EUSK en el sistema y los dos restantes seleccionaron los ítems de prueba y en el paso 2 eligieron la calibración mediante expertos como método de calibración. En cuanto al paso 3, todos escogieron la opción *No* en el campo *Administración de expertos mediante el Moodle de CALLIE*, y como valor para el parámetro *número de expertos disponibles* introdujeron *116* dejando para el resto las opciones y valores proporcionados automáticamente por la herramienta. Respecto al paso 4, todos siguieron las directrices dadas eligiendo las opciones del penúltimo paso de la tarea. En resumen, todo participante realizó su tarea correctamente y también confirmó la petición correspondiente, que pasó a CALLIE-PRO como aceptada.

En cuanto al desarrollo de la parte IRALE, al igual que en el caso anterior todos los participantes realizaron sucesivamente los cuatro pasos del menú. En los pasos 1 y 2 del menú todos los participantes seleccionaron adecuadamente el tema IRALE y la calibración mediante expertos. En el paso 3 todos definieron el tipo de administración mediante *CALLIE-MOODLE*. Obviamente, en este caso la opción en CALLIE para el diseño de cuestionarios fue la de cuestionarios no homogéneos ya que los 4 tipos de

cuestionarios diseñados tenían distinta longitud. A continuación, siguiendo con el paso 3, indicaron los parámetros específicos para diseñar estos cuestionarios no homogéneos. Los cuatro tipos de cuestionarios diseñados implicaban la división del banco de ítems en cinco bloques de ítems disjuntos, esto es 5 bloques con los distintos ítems y los 8 expertos disponibles. Durante la introducción de estos datos, CALLIE generó automáticamente dos alertas debidas al bajo número de expertos captados y a la posibilidad de no llegar a las 7 opiniones por ítem, que en ningún caso les impidieron seguir con el proceso de calibración. Acto seguido, indicaron la composición de los distintos bloques y cuestionarios en CALLIE-EXPERT. El sistema también emitió dos avisos (en naranja) referidos a la posible duración excesiva y cantidad de opiniones, utilizando una media orientativa que en ningún caso impidió seguir con el proceso de calibración. Todos los participantes realizaron la prueba correctamente hasta aquí. Para finalizar el paso 3, se debían asignar los distintos ítems a cada uno de estos bloques, con los que se crearían posteriormente los cuatro tipos de cuestionarios, idénticos a los diseñados por IRALE. Aunque era mejor definir la configuración inicial con un reparto de ítems *continuo* solamente dos de los alumnos se percataron y el resto dejó el reparto *alterno* que está por defecto y fruto de esta elección uno de los alumnos fue incapaz de realizar esta tarea correctamente, fallando en la distribución de los ítems en los bloques puesto que no los redistribuyó acorde a la tabla proporcionada. En el paso 4 todos los participantes volvieron a acertar considerando 11 niveles posibles de dificultad y eligiendo para el resto de datos las opciones descritas en el penúltimo paso de la tarea. Todos confirmaron la petición correspondiente.

Al finalizar los 25 minutos de la prueba todos los participantes rellenaron la encuesta correspondiente. En cuanto a la puntuación sobre el menú dada por los cinco alumnos en orden fue: 8, 8, 9, 9 y 9. Media: 8,6. Los cinco participantes consideran que el componente es adecuado para su función y que funciona correctamente. Hubo dos comentarios positivos: sobre el cálculo automático de datos y las alertas sobre posibles errores. También hubo un comentario negativo: un participante consideró escasa la ayuda de las pantallas para la división en bloques.

PRUEBA 3

Preparación de un libro Excel con datos recogidos para PRO-C

La prueba consiste en preparar un libro Excel con los datos recogidos en la administración de ítems de IRALE e importarlos a CALLIE-EXPERT mediante PRO-C. Los pasos a dar fueron: (1) transformar un fichero Excel llamado datos_admin.xlsx, que contiene todos los datos necesarios de las administraciones, en un fichero Excel con el formato adecuado para CALLIE-EXPERT, y (2) importarlo a CALLIE desde la petición correspondiente en CALLIE-PRO.

Como tarea previa la desarrolladora del experimento había creado una petición con las características del experimento ideado por IRALE, llamada *REQE_IRALE_PRUE*, en la que los datos de las administraciones se recogen desde Excel. Había creado también el fichero Excel *datos_admin.xlsx* a partir de los datos del informe generado por los de IRALE en la prueba 3 del experimento 1. En este fichero se habían añadido todos los datos que necesita el libro Excel a preparar, pero las columnas aparecían desordenadas y los niveles de dificultad estaban en formato CEFR.

Para completar esta prueba se estableció un tiempo máximo de 15 minutos.

Desarrollo de la prueba

Para crear adecuadamente el fichero Excel, los participantes debían tener en cuenta que la ayuda sobre el formato, nombre y ubicación del libro Excel a crear aparece en la opción Ver Resumen de CALLIE-PRO correspondiente a la petición de calibración y que CALLIE-EXPERT solo admite valores enteros positivos para indicar las escalas de nivel de dificultad, con lo que las estimaciones de nivel del fichero proporcionado se debían transformar a una *escala numérica uniforme de 1 a 11* correspondiente a cada nivel del CEFR, es decir, se asignar 1 a A1, 2 a A1+ y así sucesivamente hasta 11 a C2.

Nivel CEFR	A1	A1+	A2	A2+	B1	B1+	B2	B2+	C1	C1+	C2
Nivel CALLIE-EXPERT	1	2	3	4	5	6	7	8	9	10	11

Tabla 34 – Correspondencias entre los 11 niveles del CEFR y los niveles para CALLIE-EXPERT.

Lo anterior implica que en esta prueba el participante debía encontrar la ayuda, modificar adecuadamente la escala, ordenar las columnas, cambiar el nombre del fichero y ubicarlo en el directorio necesario. Solamente si hacía todo correctamente podría recoger bien los datos. El alumno 5 no encontró la ayuda, así que tampoco modificó la escala ni cumplió con el formato aunque sí puso bien el nombre y la ubicación. El alumno 2 hizo todo bien excepto el cambio de las escalas por lo que tampoco pudo recoger los datos. Todos los demás realizaron la prueba correctamente y en tiempo.

Al finalizar los 15 minutos de la prueba todos los participantes rellenaron la encuesta correspondiente. En cuanto a la puntuación sobre la preparación del fichero Excel y la posterior recogida de datos dada por los cinco alumnos en orden fue: 7, 6, 6, 7 y 6. Media: 6,4. Los cinco participantes consideran que el componente es adecuado para su función y que funciona correctamente. Los participantes comentaron que no hay posibilidad ni de elegir el directorio ni de poner cualquier nombre al fichero, y que la ayuda no se encuentra.

PRUEBA 4

Recogida de datos, análisis y calibración con PRO-C

Esta prueba se divide en dos partes que corresponden a las calibraciones del banco de Hezinet y el del banco de IRALE respectivamente.

La **parte Hezinet** de la prueba consistió en analizar, calibrar y obtener los resultados detallados de una petición de calibración idéntica a la que correspondería al experimento de calibración llevado a cabo manualmente para Hezinet. Los pasos a dar fueron: (1) recoger los datos de las administraciones desde el fichero Excel proporcionado, (2) analizar y calibrar, y (3) ver los resultados.

Como tarea previa para la parte Hezinet la desarrolladora del experimento ya había realizado la prueba 2 con CALLIE-EXPERT creando una petición con código REQE_EUSK_20130501_121159. También había creado un libro Excel con los datos recogidos en las administraciones de ítems de la prueba original de Arruabarrena (2010). Estos datos se encontraban en la base de datos del experimento original, en la muestra que se había obtenido originalmente mediante las pruebas de campo y que

constaba de 4887 entradas correspondientes a las aportaciones de los 116 expertos participantes sobre un banco de 252 ítems. Estas 4887 aportaciones se importaron a un fichero Excel que alimentaría posteriormente al sistema de calibración CALLIE. Concretamente se utilizó de cada entrada el identificativo del experto, el identificativo del ítem, el identificador del cuestionario, la respuesta dada por el experto al ítem que además se codificó adecuadamente para CALLIE (correcta, incorrecta o sin contestar) y el nivel estimado por el experto para ese ítem. Por último, cuando había varios niveles en la misma entrada se separaron en varias filas. Como resultado se obtuvo una nueva muestra de 5100 entradas, con los identificativos de los expertos participantes y los identificativos de los 252 ítems renombrados como EUSKX para compatibilizarlos con los identificativos del banco EUSK utilizado en el sistema CALLIE-EXPERT.

La **parte IRALE** de la prueba consistió en replicar el análisis y calibración de IRALE, esto es la prueba 4 del experimento 1, incluida la recogida de los datos desde Moodle. Para ello, se pidió a cada participante que lleve a cabo dos casos con CALLIE-EXPERT sobre los mismos datos. Caso 1: aplicar el análisis que utiliza *la malla más restrictiva* recomendada para el criterio C.it-2 (85% y 3 niveles horquilla), que es el supuesto inicial que se introdujo en la prueba 2 y, a la luz de los resultados obtenidos, responder a cuatro preguntas: 1. cantidad de ítems con la dificultad calculada, 2. total de opiniones utilizadas en los cálculos de la dificultad, 3. criterio que ha descartado algún ítem junto a la cantidad de ítems descartados por ese criterio, y 4. parámetros del diseño que han descartado algún ítem. Caso 2: *Simular resultados alternativos* modificando los parámetros anteriores hasta obtener valores de dificultad para todos los ítems del banco.

Como tarea previa para la parte IRALE la desarrolladora del experimento ya había realizado la prueba 2 con CALLIE-EXPERT creando una petición con código REQE_IRALE_20130820_112134. También se había realizado ya el experimento 1 y los datos aportados por los expertos estaban almacenados y revisados en el curso correspondiente de CALLIE-MOODLE que se asoció a la petición anterior.

Para completar esta prueba se estableció un tiempo máximo de 30 minutos.

Desarrollo de la prueba

En cuanto al desarrollo de la parte Hezinet, los cinco alumnos realizaron sucesivamente la recogida de datos desde Excel, el análisis y la obtención y visualización de los resultados mediante las opciones correspondientes de la interfaz CALLIE-PRO.

En cuanto al desarrollo de la parte IRALE, al igual que en el caso anterior todos los participantes realizaron sucesivamente la recogida de datos, el análisis y la obtención y visualización de los resultados mediante CALLIE-PRO. Todos los participantes respondieron adecuadamente a las preguntas: ítems con dificultad calculada (80) el resto de ítems (51) fueron descartados por el criterio C.it-2, debido a los parámetros porcentaje de valoraciones (85%) y niveles (3) en horquilla. Y total de opiniones utilizadas (420) como suma de la columna TotalVals en la hoja CalculosDificultad. Hubo diferencias en la manera de llegar a las respuestas: mientras que para responder a la cantidad de ítems calibrados dos de ellos utilizaron el propio resumen de resultados contando manualmente, los otros 3 emplearon la hoja Excel CalculosDificultad. También hubo diferencias en el modo de obtener los parámetros, ya que cuatro utilizaron el resumen de resultados y uno la opción Ver resumen de la

pantalla de CALLIE-PRO con la lista de calibraciones. A continuación, se llevó a cabo el caso 2, en el que todos los participantes obtuvieron varios Excel aumentando la horquilla (de 3 a 4) y/o bajando el porcentaje de valoraciones (menos del 85%) y llegaron a un resultado satisfactorio, como se refleja en la tabla del final de este anexo.

En resumen, todo participante realizó sus tareas correctamente y dentro del tiempo establecido para la prueba.

Al finalizar los 30 minutos de la prueba todos los participantes rellenaron la encuesta correspondiente. En este caso se pidió a cada participante dos puntuaciones, una sobre el control del proceso con PRO-C y otra sobre su utilidad para discutir los resultados obtenidos.

En cuanto a la puntuación sobre el progreso del proceso dada por los cinco alumnos en orden fue: 9, 9, 9, 10 y 8. Media: 9. Los cinco participantes consideran que el componente es adecuado para su función y que funciona correctamente. Hubo varios comentarios positivos: sobre el hecho de que el sistema sea capaz de realizar automáticamente todo el proceso, y con mensajes que explican cada paso a dar. Como comentario negativo se indicó que la lista de peticiones no se puede ordenar.

En cuanto a la puntuación sobre la discusión de resultados dada por los cinco alumnos en orden fue: 8, 9, 8, 8 y 8. Media: 8,2. Los cinco participantes consideran que el componente es adecuado para su función y que funciona correctamente. Hubo dos comentarios positivos: los resultados se pueden ordenar y la información que se da es clara.

RESULTADOS PARA LAS RÉPLICAS DE LAS CALIBRACIONES DE LOS BANCOS

La utilización de los resultados intermedios conseguidos en el experimento 1 con los miembros de IRALE, junto a la preparación previa y celebración sucesiva de las pruebas 2 y 4 de este experimento 2 crean una serie de resultados en CALLIE-EXPERT que permiten replicar con la herramienta las calibraciones del banco de ítems de Hezinet y del banco de ítems de IRALE.

Concretamente, como **resultado de la prueba 4** se pueden obtener con CALLIE-EXPERT las réplicas de las calibraciones en dificultad sobre los dos bancos de ítems: la del banco de ítems Hezinet como resultado del procesado de la petición con código REQE_EUSK_20130501_121159 y la del banco de ítems IRALE como resultado del procesado inicial de la petición con código REQE_IRALE_20130820_112134 y su posterior simulación con 3 niveles y una malla del 60%.

REVISIÓN DE LAS TAREAS INVOLUCRADAS

Para llevar a cabo la evaluación de cada componente de CALLIE-EXPERT se detectaron las tareas involucradas en cada prueba y se realizó una revisión de las mismas.

El resultado de esta revisión se resume en la tabla siguiente, en la que se refleja la participación de los cinco alumnos en las distintas tareas y el tiempo total consumido en cada prueba. En cada tarea aparece si el participante la realizó adecuadamente o no (Ok o Mal), y en la última columna aparece el grado de realización (o porcentaje global de participantes con realización correcta).

	Tareas involucradas	Alumno1	Alumno2	Alumno3	Alumno4	Alumno5	Grado de realización
Prueba 1	**Selección banco**	Ok	Mal	Ok	Ok	Ok	**80%**
	Introducción mediante utilidad	Ok	Ok	Ok	Ok	Ok	**100%**
	Introducción mediante importación	Ok	Ok	Ok	Ok	Mal	**80%**
	TIEMPO TOTAL prueba 1 (máximo 15 minutos)	**9'38''**	**10'25''**	**9'20''**	**12'53''**	**15'**	**11'03''**
Prueba 2 (parte Hezinet)	**Paso 1**	Sel. banco	Sel. banco	Ítems prueba	Sel. banco	Ítems prueba	**100%**
	Paso 2	Ok	Ok	Ok	Ok	Ok	**100%**
	Paso 3	Ok	Ok	Ok	Ok	Ok	**100%**
	Paso 4	Ok	Ok	Ok	Ok	Ok	**100%**
	Confirmar Petición	Ok	Ok	Ok	Ok	Ok	**100%**
Prueba 2 (parte IRALE)	**Pasos 1 y 2**	Ok	Ok	Ok	Ok	Ok	**100%**
	Reparto inicial	Continuo	Alterno	Continuo	Alterno	Alterno	**40%**
	Resto paso 3	Ok	Ok	Ok	Ok	Mal	**80%**
	Paso 4	Ok	Ok	Ok	Ok	Ok	**100%**
	Confirmar Petición	Ok	Ok	Ok	Ok	Ok	**100%**
	TIEMPO TOTAL prueba 2 (máximo 25 minutos)	**20'58''**	**24'30''**	**18'22''**	**23'45''**	**13'09''**	**20'06''**
Prueba 3	**Encontrar la ayuda**	Ok	Ok	Ok	Ok	Mal	**80%**
	Modificar escala	Ok	Mal	Ok	Ok	Mal	**60%**
	Reordenar columnas	Ok	Ok	Ok	Ok	Mal	**80%**
	Nombre y ubicación	Ok	Ok	Ok	Ok	Ok	**100%**
	Recoger datos desde Excel a la petición de prueba	Ok	Mal	Ok	Ok	Mal	**60%**
	TIEMPO TOTAL prueba 3 (máximo 15 minutos)	**10'32''**	**15'**	**8'17''**	**9'03''**	**15'**	**11'30''**
Prueba 4 (parte Hezinet)	**Recoger datos desde Excel**	Ok	Ok	Ok	Ok	Ok	**100%**
	Depurar y calibrar	Ok	Ok	Ok	Ok	Ok	**100%**
	Ver resultados detallados	Ok	Ok	Ok	Ok	Ok	**100%**
Prueba 4 (parte IRALE)	**Recoger datos desde Moodle**	Ok	Ok	Ok	Ok	Ok	**100%**
	Depurar y calibrar	Ok	Ok	Ok	Ok	Ok	**100%**
	Ver resultados detallados	Ok	Ok	Ok	Ok	Ok	**100%**
	Responder preguntas	Ok	Ok	Ok	Ok	Ok	**100%**
	Simular (% aplicado, niveles)	60%,3	60%,4	50%,3	55%,3	65%,3	**100%**
	TIEMPO TOTAL prueba 4 (máximo 30 minutos)	**24'00''**	**27'45''**	**30'**	**28'01''**	**25'55''**	**26'8''**

Tabla 35 – Revisión de tiempos y tareas de los participantes en el experimento 2.

A4 Informe de resultados réplica Hezinet

En este anexo se presenta el informe de resultados con los detalles relevantes obtenidos por CALLIE-EXPERT durante la calibración del banco de ítems para el caso Hezinet. Concretamente, los datos que aparecen en este anexo se han transcrito directamente desde el libro Excel que genera el sistema en las hojas AdministracionesCuestionarioExp (Tabla 36), CalculosDificultad (Tabla 37 y Tabla 38) e ItemsIntentados (Tabla 39) descritas en el capítulo X de esta memoria. Por otro lado, los enunciados de los 252 ítems originales y los datos recogidos se pueden encontrar en Arruabarrena (2010).

En el primer y el segundo campo de la hoja AdministracionesCuestionarioExp se determina la administración mediante el identificador del cuestionario y del experto respectivamente (columnas *IDcuest* e *IDexperto*), en los tres siguientes se recogen los valores necesarios para calcular la tasa de acierto del experto (columnas *RespsExperto*, *RespsCorrectas* y *%Acierto*) y en el último campo se indican los motivos del rechazo en las administraciones que fueron rechazadas por CALLIE-EXPERT.

IDcuest	IDexperto	Resps Experto	Resps Correctas	%Acierto	Razón si se rechaza
7	**1**	33	33	100	
8	**2**	32	29	90,62	
6	**3**	31	28	90,32	
5	**4**	30	30	100	
4	**5**	32	32	100	
3	**6**	28	26	92,86	
2	**7**	30	29	96,67	
1	**8**	29	29	100	
6	**9**	34	34	100	
7	**10**	33	32	96,97	
8	**11**	32	30	93,75	
8	**12**	30	29	96,67	
6	**13**	33	33	100	
5	**14**	21	18	85,71	
4	**15**	30	29	96,67	
6	**16**	34	34	100	
8	**17**	26	26	100	
7	**18**	32	31	96,88	
8	**19**	31	30	96,77	
5	**20**	29	27	93,1	
5	**21**	30	29	96,67	

IDcuest	IDexperto	Resps Experto	Resps Correctas	%Acierto	Razón si se rechaza
4	**22**	32	29	90,62	
3	**23**	32	32	100	
1	**24**	29	29	100	
2	**25**	31	31	100	
2	**26**	0	0	0	C.ex-2(1): Tasa de acierto PROPIO no superior al 75%
3	**27**	32	29	90,62	
1	**28**	29	29	100	
1	**29**	9	9	100	
2	**30**	25	25	100	
3	**31**	32	31	96,88	
4	**32**	26	26	100	
5	**33**	29	28	96,55	
6	**34**	32	30	93,75	
1	**35**	29	29	100	
2	**36**	32	32	100	
3	**37**	32	31	96,88	
4	**38**	31	30	96,77	
6	**39**	9	9	100	
8	**40**	31	30	96,77	
3	**41**	11	11	100	
2	**42**	31	31	100	
1	**43**	13	12	92,31	
7	**44**	32	26	81,25	
1	**45**	29	29	100	
2	**46**	30	30	100	
3	**47**	32	29	90,62	
4	**48**	29	29	100	
5	**49**	30	29	96,67	
6	**50**	34	34	100	
7	**51**	33	33	100	
8	**52**	32	31	96,88	
1	**53**	29	29	100	
2	**54**	32	32	100	
4	**55**	32	29	90,62	
8	**56**	31	30	96,77	
1	**57**	28	28	100	
2	**58**	32	32	100	
3	**59**	32	32	100	
4	**60**	32	32	100	
5	**61**	30	30	100	
20	**80**	30	30	100	
22	**81**	34	31	91,18	

IDcuest	IDexperto	Resps Experto	Resps Correctas	%Acierto	Razón si se rechaza
24	**82**	30	30	100	
21	**83**	30	28	93,33	
23	**84**	34	31	91,18	
25	**85**	31	27	87,1	
26	**86**	31	29	93,55	
27	**87**	32	29	90,62	
29	**88**	28	27	96,43	
30	**89**	30	29	96,67	
31	**90**	28	28	100	
32	**91**	0	0	0	C.ex-2(1): Tasa de acierto PROPIO no superior al 75%
33	**92**	30	29	96,67	
10	**201**	29	29	100	
9	**202**	26	26	100	
14	**203**	34	34	100	
13	**204**	28	27	96,43	
10	**205**	30	29	96,67	
9	**206**	27	27	100	
12	**207**	30	28	93,33	
11	**208**	34	32	94,12	
14	**209**	31	31	100	
13	**210**	28	28	100	
10	**211**	30	29	96,67	
9	**212**	27	27	100	
14	**213**	32	31	96,88	
13	**214**	29	29	100	
12	**215**	29	27	93,1	
11	**216**	34	33	97,06	
10	**217**	30	29	96,67	
9	**218**	26	25	96,15	
13	**219**	27	26	96,3	
12	**220**	28	27	96,43	
11	**221**	34	32	94,12	
11	**222**	34	33	97,06	
10	**223**	30	30	100	
10	**224**	30	28	93,33	
11	**225**	34	33	97,06	
11	**226**	34	34	100	
10	**227**	30	30	100	
13	**228**	28	27	96,43	
12	**229**	28	25	89,29	
9	**230**	0	0	0	C.ex-2(1): Tasa de acierto PROPIO no superior al 75%
14	**231**	0	0	0	C.ex-2(1): Tasa de acierto PROPIO no superior al 75%

IDcuest	IDexperto	Resps Experto	Resps Correctas	%Acierto	Razón si se rechaza
13	**232**	24	21	87,5	
12	**233**	30	26	86,67	
10	**234**	0	0	0	C.ex-1: No estima adecuadamente el nivel en ningún ítem
9	**235**	25	22	88	
12	**236**	30	30	100	
14	**237**	36	36	100	
13	**238**	28	27	96,43	
10	**239**	29	29	100	
9	**240**	26	26	100	
13	**241**	6	6	100	
14	**242**	6	6	100	

Tabla 36 – Detalle de los resultados para los expertos en el caso Hezinet.

Las dos tablas siguientes se han obtenido directamente de la hoja CalculosDificultad del informe de resultados. Esta hoja guarda las frecuencias relativas de los pronósticos otorgadas por los expertos sobre la dificultad de los ítems (columna *FrecuenciasEstimadas*), es decir, las frecuencias recopiladas para cada nivel a estimar. Además, almacena las características de los ítems de Hezinet que presentaron intervalos ambiguos a la hora de calcular la dificultad.

La siguiente tabla recoge las frecuencias relativas de los pronósticos otorgadas por los expertos sobre la dificultad de los ítems (columnas *1-12*). 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 ítems ambiguos y los juicios considerados por M.dif-2.

Nivel Item#	1	2	3	4	5	6	7	8	9	10	11	12	TotalVals
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
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

Nivel / Item#	1	2	3	4	5	6	7	8	9	10	11	12	TotalVals
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
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
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

Nivel Item#	1	2	3	4	5	6	7	8	9	10	11	12	TotalVals
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
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

Nivel Item#	1	2	3	4	5	6	7	8	9	10	11	12	TotalVals
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
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
193				2	3	4	4	1		2			16
195			1	2	3	3			1				10
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

Nivel Item#	1	2	3	4	5	6	7	8	9	10	11	12	TotalVals
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 37 – Frecuencias utilizadas en el cálculo de la dificultad en el caso Hezinet.

En la siguiente tabla la columna *ValsIni* guarda la máxima frecuencia encontrada en los niveles consecutivos indicados para la horquilla y que cumple el porcentaje en horquilla requerido. *Vals_X* son las valoraciones encontradas al añadir un nivel más en la horquilla (es decir, añadiendo el siguiente nivel a *Inter_X*), *DesvT_X* contiene el valor de la desviación típica para ese nuevo intervalo que contiene un nivel adicional. *Interv* contiene el intervalo ganador sobre el que se realiza el cálculo de la media. Por último, la columna *Dificultad* contiene la estimación obtenida para la dificultad, que está entre 1 y el número de niveles.

Item#	ValsIni	Inter_1	V_1	DesvT_1	Inter_2	V_2	DesvT_2	Inter_3	V_3	Interv	Dificultad
20	14	3-6	15	1,0456	4-7	15	1,2038			**3-7**	**4,8**
35	11	2-5	12		3-6	13	1,25	4-7	11	**3-7**	**4,7692**
49	12	2-5	13	1,141	3-6	12				**2-6**	**4,0769**
72	61	3-6	66	0,998	4-7	61				**3-7**	**4,6061**
76	14	2-5	15	0,8536	4-6	14				**2-6**	**4,2667**
82	10	2-5	11	1,0833	3-6	10				**2-6**	**4,0909**
87	14	3-6	15	0,9798	4-7	14				**3-7**	**4,8**
91	13	4-7	15	1,1924	5-8	14				**4-8**	**5,6667**
99	11	4-7	12	1,1428	5-8	12	1,299			**4-8**	**5,8333**
107	10	4-7	12	1,3122	5-8	11				**4-8**	**5,6667**
126	13	6-9	14	1,0994	7-10	13				**6-10**	**7,9286**
159	14	5-8	15	0,9571	6-9	15	1,0239			**5-9**	**7,1333**
166	14	7-10	16	1,2484	8-11	14				**7-11**	**9,0625**

Item#	ValsIni	Inter_1	V_1	DesvT_1	Inter_2	V_2	DesvT_2	Inter_3	V_3	Interv	Dificultad
167	12	5-8	13		6-9	14	1,2935			**6-10**	**7,4286**
171	16	5-8	17	0,8065	7-9	16				**5-9**	**7,2353**
185	13	4-7	14	0,9895	5-8	14	1,0522			**4-8**	**6,1429**
198	13	3-6	14	1,125	4-7	13				**3-7**	**4,8571**
215	13	6-9	15	1,1924	7-10	13				**6-10**	**7,6667**
222	13	6-9	15	1,2035	7-10	13				**6-10**	**7,8667**
231	12	7-10	13		8-11	14	1,5203			**8-12**	**9,2143**

Tabla 38 – Cálculo detallado de la dificultad en los 20 ítems ambiguos en el caso Hezinet.

En la hoja ItemsIntentados se indica el porcentaje de respuestas correctas obtenidas para cada ítem y el total de valoraciones utilizadas en el cálculo de la dificultad. En el último campo se indica bien el valor obtenido para la dificultad de ese ítem o bien el motivo de eliminación – C.it-1 o C.it-2 – indicando entre paréntesis la iteración en la que se produjo junto con su mensaje aclaratorio.

Item#	IDCallie	PorcResp Correcta	totalVals	Dificultad/Razón eliminación
1	EUSK1	93,33	14	1,1429
2	EUSK2	94,74	18	1,75
3	EUSK3	94,12	17	1,6471
4	EUSK4	81,25	16	1,7857
5	EUSK5	94,12	16	1,6875
6	EUSK6	92,86	13	1,4615
7	EUSK7	93,75	63	1,6885
8	EUSK8	100	18	2,125
9	EUSK9	100	16	2,0714
10	EUSK10	86,67		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
11	EUSK11	68,75		C.it-2(2): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
12	EUSK12	100	14	2,1818
13	EUSK13	100	16	2,3077
14	EUSK14	93,33	14	4,5385
15	EUSK15	94,44	17	1,375
16	EUSK16	46,67		C.it-1: El porcentaje de respuestas correctas (46,67%) no llega al 50%
17	EUSK17	87,5	15	4,3571
18	EUSK18	84,62	12	1,7273
19	EUSK19	86,67	14	1,7692
20	EUSK20	68,75	16	4,8
21	EUSK21	100	14	1,8571
22	EUSK22	93,75	15	2,5
23	EUSK23	92,86	13	2
24	EUSK24	93,75		C.it-2(1): Máximo valoraciones en horquilla (8) menor que el mínimo requerido (12)
25	EUSK25	5,88		C.it-1: El porcentaje de respuestas correctas (5,88%) no llega al 50%

Item#	IDCallie	PorcResp Correcta	totalVals	Dificultad/Razón eliminación
26	EUSK26	92,86	13	1,9091
27	EUSK27	98,28	57	1,8302
28	EUSK28	94,12	16	1,5
29	EUSK29	100	16	2
30	EUSK30	100		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
31	EUSK31	94,44	17	2,6471
32	EUSK32	93,33	14	3,5385
33	EUSK33	100	16	2,8667
34	EUSK34	100	13	4,5
35	EUSK35	93,33	14	4,7692
36	EUSK36	94,44	17	2,875
37	EUSK37	100	14	4,25
38	EUSK38	100	16	3,4286
39	EUSK39	92,86	13	1,5385
40	EUSK40	94,44	17	4,5
41	EUSK41	100	15	3,3333
42	EUSK42	100	16	3,5385
43	EUSK43	93,75		C.it-2(2): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
44	EUSK44	92,31	12	3,2222
45	EUSK45	100	18	3,5882
46	EUSK46	94,44	17	3,2
47	EUSK47	96,77	61	3,2679
48	EUSK48	93,33	14	3,5
49	EUSK49	92,86	13	4,0769
50	EUSK50	94,74		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (15)
51	EUSK51	87,5		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
52	EUSK52	92,86	13	3,0769
53	EUSK53	94,44	17	3,4375
54	EUSK54	100	17	3,4375
55	EUSK55	93,75	16	2,9333
56	EUSK56	92,86		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (11)
57	EUSK57	85,71	13	2,25
58	EUSK58	94,12	16	4,1333
59	EUSK59	94,44		C.it-2(1): Máximo valoraciones en horquilla (13) menor que el mínimo requerido (14)
60	EUSK60	93,75		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
61	EUSK61	88,89	17	4,7143
62	EUSK62	100		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
63	EUSK63	93,75		C.it-2(1): Máximo valoraciones en horquilla (9) menor que el mínimo requerido (12)

Item#	IDCallie	PorcResp Correcta	totalVals	Dificultad/Razón eliminación
64	EUSK64	93,33	14	3,3333
65	EUSK65	94,44	17	3,6875
66	EUSK66	100	15	3
67	EUSK67	100	16	4,5714
68	EUSK68	93,75	15	4,7692
69	EUSK69	93,33	14	4,9231
70	EUSK70	100	16	4,6
71	EUSK71	100	15	4,4
72	EUSK72	97,01	66	4,6061
73	EUSK73	94,12	16	4,625
74	EUSK74	92,86	13	4,5
75	EUSK75	93,75	16	4,8571
76	EUSK76	100	15	4,2667
77	EUSK77	93,33	14	3,7857
78	EUSK78	94,44	17	4,8125
79	EUSK79	100	14	4,25
80	EUSK80	100		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
81	EUSK81	94,12	16	4,8571
82	EUSK82	92,31	12	4,0909
83	EUSK83	100		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (15)
84	EUSK84	72,22	17	5,25
85	EUSK85	93,33	14	5,1429
86	EUSK86	94,44	17	5,1538
87	EUSK87	100	15	4,8
88	EUSK88	100	15	4,3333
89	EUSK89	86,67	14	5,1538
90	EUSK90	94,44	17	4,5714
91	EUSK91	100	16	5,6667
92	EUSK92	95,16	61	4,6071
93	EUSK93	100	14	5,0714
94	EUSK94	94,12	16	5,375
95	EUSK95	92,31	12	4,8
96	EUSK96	94,12	17	7,4667
97	EUSK97	93,75	16	5,2857
98	EUSK98	94,12	16	5,5714
99	EUSK99	92,86	13	5,8333
100	EUSK100	100	17	5,25
101	EUSK101	100	16	5,1538
102	EUSK102	92,86	13	6,9167
103	EUSK103	94,44		C.it-2(1): Máximo valoraciones en horquilla (12) menor que el mínimo requerido (14)

Item#	IDCallie	PorcResp Correcta	totalVals	Dificultad/Razón eliminación
104	EUSK104	100	16	6,3077
105	EUSK105	93,33		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
106	EUSK106	94,12	16	5,0769
107	EUSK107	92,86	13	5,6667
108	EUSK108	100	16	6,0625
109	EUSK109	100	14	6,0769
110	EUSK110	93,33	14	5,2857
111	EUSK111	94,44	17	5,5714
112	EUSK112	95,24	61	5,6415
113	EUSK113	100	15	6
114	EUSK114	93,75	16	5,4286
115	EUSK115	93,33		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
116	EUSK116	94,74	18	6,3529
117	EUSK117	100	15	5,2308
118	EUSK118	100	15	5,7692
119	EUSK119	93,75	15	5,6667
120	EUSK120	92,86	13	6,0769
121	EUSK121	75	15	7
122	EUSK122	100	13	6,1538
123	EUSK123	81,25	15	6,3333
124	EUSK124	92,31	12	5,5833
125	EUSK125	93,75	15	6,1538
126	EUSK126	93,75	15	7,9286
127	EUSK127	93,33	14	6,8333
128	EUSK128	94,44	17	6,8667
129	EUSK129	100	15	4,8462
130	EUSK130	100	16	6
131	EUSK131	93,75		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
132	EUSK132	93,33	14	8,3333
133	EUSK133	100	16	5,3125
134	EUSK134	93,75		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
135	EUSK135	86,67		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
136	EUSK136	93,75		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
137	EUSK137	93,55	61	3,5
138	EUSK138	87,5	16	8,3125
139	EUSK139	93,75	16	9,5
140	EUSK140	93,33	14	6,4167
141	EUSK141	87,5	15	6,3077
142	EUSK142	100	16	5,7143

Item#	IDCallie	PorcResp Correcta	totalVals	Dificultad/Razón eliminación
143	EUSK143	100		C.it-2(1): Máximo valoraciones en horquilla (12) menor que el mínimo requerido (13)
144	EUSK144	88,24		C.it-2(1): Máximo valoraciones en horquilla (12) menor que el mínimo requerido (13)
145	EUSK145	78,57		C.it-2(1): Máximo valoraciones en horquilla (9) menor que el mínimo requerido (11)
146	EUSK146	93,75	15	5,1667
147	EUSK147	94,12	16	7,2857
148	EUSK148	75		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
149	EUSK149	92,86	13	6,5
150	EUSK150	93,75	16	8
151	EUSK151	0		C.it-1: El porcentaje de respuestas correctas (0%) no llega al 50%
152	EUSK152	92,86	13	7,3636
153	EUSK153	87,5		C.it-2(2): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
154	EUSK154	100	15	10,7692
155	EUSK155	93,33	15	6,4615
156	EUSK156	88,24	16	7,0667
157	EUSK157	88,71		C.it-2(1): Máximo valoraciones en horquilla (40) menor que el mínimo requerido (47)
158	EUSK158	92,86	13	10,75
159	EUSK159	100	18	7,1333
160	EUSK160	93,75	15	11,1333
161	EUSK161	86,67		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
162	EUSK162	88,89	17	6,0667
163	EUSK163	100		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
164	EUSK164	100	14	7,3636
165	EUSK165	92,31	12	8
166	EUSK166	88,24	16	9,0625
167	EUSK167	100	16	7,4286
168	EUSK168	100		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
169	EUSK169	68,75		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
170	EUSK170	92,86		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (11)
171	EUSK171	100	17	7,2353
172	EUSK172	94,12	16	4,5385
173	EUSK173	88,89	17	7,3846
174	EUSK174	92,86	13	6,7
175	EUSK175	100	18	6,3333
176	EUSK176	98,41		C.it-2(1): Máximo valoraciones en horquilla (40) menor que el mínimo requerido (48)
177	EUSK177	100	15	9,5385

Item#	IDCallie	PorcResp Correcta	totalVals	Dificultad/Razón eliminación
178	EUSK178	93,75		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
179	EUSK179	94,12	16	7,1429
180	EUSK180	94,12	17	10,2143
181	EUSK181	94,44	18	5,7143
182	EUSK182	75	15	5,6429
183	EUSK183	92,86	13	6,7273
184	EUSK184	100	13	6,7
185	EUSK185	100	16	6,1429
186	EUSK186	92,86	13	5,2727
187	EUSK187	77,78	17	7,1333
188	EUSK188	100		C.it-2(1): Máximo valoraciones en horquilla (8) menor que el mínimo requerido (12)
189	EUSK189	87,5	16	9,3333
190	EUSK190	87,5		C.it-2(2): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
191	EUSK191	68,75		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
192	EUSK192	100		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
193	EUSK193	100	16	5,7692
194	EUSK194	87,5		C.it-2(2): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
195	EUSK195	90,91	10	4,8889
196	EUSK196	86,67	14	8,75
197	EUSK197	94,12	16	5,8667
198	EUSK198	93,75	15	4,8571
199	EUSK199	71,43		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (11)
200	EUSK200	84,62		C.it-2(1): Máximo valoraciones en horquilla (9) menor que el mínimo requerido (10)
201	EUSK201	93,44	59	10,5333
202	EUSK202	94,12		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (13)
203	EUSK203	93,33		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
204	EUSK204	93,75		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
205	EUSK205	47,06		C.it-1: El porcentaje de respuestas correctas (47,06%) no llega al 50%
206	EUSK206	100	13	5,3636
207	EUSK207	88,89		C.it-2(1): Máximo valoraciones en horquilla (9) menor que el mínimo requerido (14)
208	EUSK208	53,85		C.it-2(1): Máximo valoraciones en horquilla (9) menor que el mínimo requerido (10)
209	EUSK209	93,75		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
210	EUSK210	100	13	6,3
211	EUSK211	85,71		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (11)
212	EUSK212	94,12	16	8,8571

Item#	IDCallie	PorcResp Correcta	totalVals	Dificultad/Razón eliminación
213	EUSK213	71,43	14	11,3333
214	EUSK214	92,86	14	8,4545
215	EUSK215	93,75	15	7,6667
216	EUSK216	88,89	17	7,6667
217	EUSK217	88,24	17	5,7143
218	EUSK218	100		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
219	EUSK219	81,25	15	8,3333
220	EUSK220	85,71	13	5,5833
221	EUSK221	96,88	63	4,6471
222	EUSK222	75	15	7,8667
223	EUSK223	93,75	15	8
224	EUSK224	87,5	15	8,6154
225	EUSK225	78,57		C.it-2(1): Máximo valoraciones en horquilla (8) menor que el mínimo requerido (11)
226	EUSK226	66,67	15	8,5833
227	EUSK227	60	15	11,6667
228	EUSK228	87,5	15	6,3571
229	EUSK229	93,75	15	7,5833
230	EUSK230	86,67		C.it-2(1): Máximo valoraciones en horquilla (11) menor que el mínimo requerido (12)
231	EUSK231	100	16	9,2143
232	EUSK232	93,75		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
233	EUSK233	85,71		C.it-2(1): Máximo valoraciones en horquilla (8) menor que el mínimo requerido (11)
234	EUSK234	38,89		C.it-1: El porcentaje de respuestas correctas (38,89%) no llega al 50%
235	EUSK235	88,24		C.it-2(1): Máximo valoraciones en horquilla (12) menor que el mínimo requerido (13)
236	EUSK236	53,33		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
237	EUSK237	88,89		C.it-2(1): Máximo valoraciones en horquilla (13) menor que el mínimo requerido (14)
238	EUSK238	86,67	15	11,0769
239	EUSK239	25		C.it-1: El porcentaje de respuestas correctas (25%) no llega al 50%
240	EUSK240	71,43		C.it-2(1): Máximo valoraciones en horquilla (8) menor que el mínimo requerido (11)
241	EUSK241	93,33	59	6,587
242	EUSK242	75		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
243	EUSK243	100		C.it-2(1): Máximo valoraciones en horquilla (7) menor que el mínimo requerido (12)
244	EUSK244	93,33	15	9,3846
245	EUSK245	88,24	16	6,75
246	EUSK246	71,43	13	6
247	EUSK247	94,74		C.it-2(1): Máximo valoraciones en horquilla (14) menor que el mínimo requerido (15)

Item#	IDCallie	PorcResp Correcta	totalVals	Dificultad/Razón eliminación
248	EUSK248	100		C.it-2(1): Máximo valoraciones en horquilla (10) menor que el mínimo requerido (12)
249	EUSK249	37,5		C.it-1: El porcentaje de respuestas correctas (37,5%) no llega al 50%
250	EUSK250	66,67	11	7,1111
251	EUSK251	93,33	14	2,9167
252	EUSK252	82,35	16	6,9231

Tabla 39 – Detalle de los resultados de calibración para los 252 ítems de Hezinet.

A5 Informe de resultados réplica IRALE

En este anexo se presenta el informe de resultados con los detalles relevantes obtenidos por CALLIE-EXPERT durante la calibración del banco de ítems para el caso IRALE. Concretamente, los datos que aparecen en las demás se han transcrito directamente desde el libro Excel que genera el sistema en las hojas AdministracionesCuestionarioExp (Tabla 40), CalculosDificultad (Tabla 41, Tabla 42, Tabla 44 y Tabla 45) e ItemsIntentados (Tabla 43) descritas en el capítulo X de esta memoria. No se han incluido los enunciados de los 132 ítems originales ni los datos recogidos a los expertos por ser datos confidenciales de IRALE.

En el primer y el segundo campo de la hoja AdministracionesCuestionarioExp se determina la administración mediante el identificador del cuestionario y del experto respectivamente (columnas *IDcuest* e *IDexperto*), en los tres siguientes se recogen los valores necesarios para calcular la tasa de acierto del experto (columnas *RespsExperto*, *RespsCorrectas* y *%Acierto*) y en el último campo se indican los motivos del rechazo en las administraciones que fueron rechazadas por CALLIE-EXPERT.

IDcuest	IDexperto	Resps Experto	Resps Correctas	%Acierto	Razón si se rechaza
Todos	**exp1**	132	132	100	
Todos	**exp2**	132	132	100	
Bajo	**exp3**	82	75	88,24	
Alto.85	**exp4**	80	77	90,59	
Alto.67	**exp5**	67	67	100	
Alto.67	**exp6**	62	57	85,07	
Alto.85	**exp7**	85	71	83,53	
Bajo	**exp8**	74	56	65,88	No se eliminan administraciones

Tabla 40 – Detalle de los resultados para los expertos en IRALE.

Se llevaron a cabo dos casos con CALLIE-EXPERT sobre los mismos datos. Caso 1: aplicar el análisis que utiliza *la malla más restrictiva* recomendada para el criterio C.it-2 (85% y 3 niveles horquilla) y Caso 2: *relajación de la malla al 60%* (60% y 3 niveles horquilla).

En las tres tablas siguientes se detallan los resultados que se obtuvieron con CALLIE-EXPERT en el Caso 1. 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 ítems ambiguos y los juicios considerados por M.dif-2. Además, en la columna *TotalVals* se han resaltado las que no llegaron al mínimo recomendado de 7 juicios. En este caso existen 11 niveles posibles para la dificultad de los ítems (columnas *1-11*).

Nivel Item#	1	2	3	4	5	6	7	8	9	10	11	TotalVals
1	2	1	1	0	0	0	0	0	0	0	0	4
2	1	1	2	0	0	0	0	0	0	0	0	4
3	0	0	3	1	0	0	0	0	0	0	0	4
4	3	0	1	0	0	0	0	0	0	0	0	4
5	1	2	1	0	0	0	0	0	0	0	0	4
6	0	0	0	0	2	1	1	0	0	0	0	4
7	2	1	1	0	0	0	0	0	0	0	0	4
8	3	1	0	0	0	0	0	0	0	0	0	4
9	0	0	0	0	3	1	0	0	0	0	0	4
10	0	0	2	0	1	1	0	0	0	0	0	4
11	0	0	1	1	0	0	0	0	0	0	0	2
12	0	0	0	0	1	1	0	0	0	0	0	2
13	0	1	3	0	0	0	0	0	0	0	0	4
14	0	0	1	0	2	1	0	0	0	0	0	4
15	0	1	1	0	0	0	0	0	0	0	0	2
16	0	0	1	0	2	1	2	0	0	0	0	6
17	0	0	1	0	2	0	1	0	0	0	0	4
18	0	0	4	0	0	0	0	0	0	0	0	4
19	0	0	1	1	2	0	0	0	0	0	0	4
20	0	0	0	0	2	2	0	0	0	0	0	4
21	0	0	1	1	2	0	0	0	0	0	0	4
22	0	0	0	0	4	0	0	0	0	0	0	4
23	0	0	0	0	1	3	2	0	0	0	0	6
24	0	0	0	0	1	1	2	0	1	0	0	5
25	0	0	0	0	1	0	7	0	0	0	0	8
26	0	0	1	0	1	2	0	0	2	0	0	6
27	0	0	1	0	0	3	1	0	1	0	0	6
28	0	0	0	0	0	0	7	1	0	0	0	8
29	0	0	0	0	1	0	6	0	0	0	0	7
30	0	0	0	0	2	0	6	0	0	0	0	8
31	0	0	0	0	0	0	2	2	2	0	0	6
32	0	0	0	0	0	0	3	1	2	0	0	6
33	0	0	1	0	1	2	4	0	0	0	0	8
34	0	0	0	0	2	1	4	0	1	0	0	8
35	0	0	1	0	0	0	5	2	0	0	0	8
36	0	0	0	0	0	1	0	4	1	0	0	6
37	0	0	0	0	0	0	2	2	2	0	0	6
38	0	0	0	0	0	1	1	2	2	0	0	6
39	0	0	0	0	2	1	2	3	0	0	0	8
40	0	0	0	0	2	1	4	1	0	0	0	8

Nivel Item#	1	2	3	4	5	6	7	8	9	10	11	TotalVals
41	0	0	0	0	0	2	2	1	1	0	0	6
42	0	0	0	0	0	1	2	2	1	0	0	6
43	0	0	0	0	0	0	1	1	4	0	0	6
44	0	0	0	0	0	1	1	0	4	0	0	6
45	0	0	0	0	0	1	0	1	4	0	0	6
46	0	0	0	0	0	0	1	0	4	1	0	6
47	0	0	0	0	0	0	0	2	3	0	0	5
48	0	0	0	0	2	0	3	1	2	0	0	8
49	0	0	0	0	0	0	2	2	2	0	0	6
50	0	0	0	0	0	0	1	3	2	0	0	6
51	0	0	0	0	2	0	6	0	0	0	0	8
52	0	0	0	0	2	1	3	1	1	0	0	8
53	0	0	0	0	0	0	5	2	1	0	0	8
54	0	0	0	0	0	0	1	1	3	0	1	6
55	0	0	0	0	0	0	0	2	3	0	1	6
56	0	0	0	0	0	1	2	2	1	0	0	6
57	0	0	0	0	0	0	3	2	1	0	0	6
58	0	0	0	0	0	0	0	2	4	0	0	6
59	0	0	0	0	2	0	2	2	2	0	0	8
60	0	0	0	0	0	0	1	1	4	0	0	6
61	1	0	1	0	2	0	0	0	0	0	0	4
62	1	0	1	0	1	1	0	0	0	0	0	4
63	1	0	3	0	0	0	0	0	0	0	0	4
64	0	0	1	0	3	0	0	0	0	0	0	4
65	0	0	0	2	0	0	0	0	0	0	0	2
66	1	0	1	1	1	0	0	0	0	0	0	4
67	0	0	0	2	0	0	0	0	0	0	0	2
68	0	0	2	0	2	0	0	0	0	0	0	4
69	2	1	1	0	0	0	0	0	0	0	0	4
70	0	0	0	0	2	2	1	0	0	0	0	5
71	0	0	1	0	1	2	1	1	0	0	0	6
72	1	0	1	0	2	2	0	0	0	0	0	6
73	0	0	0	0	0	0	3	1	2	0	0	6
74	0	0	0	1	0	0	7	0	0	0	0	8
75	0	0	0	0	1	1	3	1	0	0	0	6
76	0	0	0	0	0	0	4	2	0	0	0	6
77	0	0	0	0	0	0	2	1	3	0	0	6
78	0	0	0	0	0	0	0	2	3	0	0	5
79	0	0	0	0	0	0	1	2	3	0	0	6
80	0	0	0	0	0	0	2	1	3	0	0	6

Nivel Item#	1	2	3	4	5	6	7	8	9	10	11	TotalVals
81	0	0	0	0	0	1	1	1	2	0	1	6
82	0	0	0	0	0	0	3	1	0	0	0	4
83	0	0	1	0	1	1	3	0	0	0	0	6
84	0	0	1	0	0	1	6	0	0	0	0	8
85	0	0	0	0	0	0	3	2	1	0	0	6
86	0	0	0	0	0	0	2	3	1	0	0	6
87	0	0	0	0	1	2	2	0	0	0	0	5
88	0	0	0	0	2	1	5	0	0	0	0	8
89	0	0	0	0	2	1	4	1	0	0	0	8
90	0	0	0	0	0	1	0	0	4	0	0	5
91	0	0	0	0	1	0	2	0	4	0	0	7
92	0	0	0	0	0	0	1	1	3	0	1	6
93	0	0	0	0	0	1	0	1	4	0	0	6
94	0	0	2	1	1	0	0	0	0	0	0	4
95	1	1	0	1	1	0	0	0	0	0	0	4
96	0	0	0	0	0	0	2	1	2	0	1	6
97	1	0	1	0	2	0	0	0	0	0	0	4
98	0	0	0	0	0	0	1	1	4	0	0	6
99	0	0	1	0	2	1	3	0	0	0	0	7
100	0	0	0	0	1	1	3	1	1	0	0	7
101	0	0	0	0	0	0	0	0	5	0	1	6
102	0	0	2	0	0	0	3	2	1	0	0	8
103	0	0	0	0	0	0	0	2	4	0	0	6
104	0	0	0	0	0	0	0	0	1	1	0	2
105	0	0	0	0	0	1	1	2	2	0	0	6
106	0	0	0	0	0	0	4	2	0	0	0	6
107	0	0	1	0	1	2	2	0	0	0	0	6
108	0	0	0	0	0	0	1	2	3	0	0	6
109	0	0	0	0	1	2	4	1	0	0	0	8
110	0	0	1	0	1	0	2	0	3	0	0	7
111	0	0	0	0	1	1	2	0	4	0	0	8
112	0	0	0	0	0	0	0	0	3	2	1	6
113	0	0	2	0	1	1	0	0	0	0	0	4
114	0	0	0	0	2	2	2	0	0	0	0	6
115	0	0	2	0	1	1	0	0	0	0	0	4
116	2	0	1	0	1	0	0	0	0	0	0	4
117	0	0	1	0	3	0	0	0	0	0	0	4
118	0	0	1	0	3	0	0	0	0	0	0	4
119	0	0	2	0	2	0	0	0	0	0	0	4
120	1	1	2	0	0	1	0	0	0	0	0	5

Nivel Item#	1	2	3	4	5	6	7	8	9	10	11	TotalVals
121	0	0	2	1	1	0	0	0	0	0	0	4
122	0	1	1	1	1	0	0	0	0	0	0	4
123	0	0	0	0	2	0	0	0	0	0	0	2
124	0	0	1	0	3	0	0	0	0	0	0	4
125	0	0	0	0	2	0	0	0	0	0	0	2
126	0	0	2	0	1	1	2	0	0	0	0	6
127	0	0	2	0	2	0	0	0	0	0	0	4
128	0	0	1	0	1	3	1	0	0	0	0	6
129	0	0	1	0	2	0	0	0	0	0	0	3
130	0	0	2	1	1	1	0	0	0	0	0	5
131	0	1	3	0	0	0	1	0	0	0	0	5
132	0	2	2	0	0	0	1	0	0	0	0	5

Tabla 41 – Frecuencias utilizadas en el cálculo de la dificultad en el caso 1 de IRALE.

La siguiente tabla refleja el cálculo para el único ítem ambiguo del Caso 1.

Item	ValsIni	Inter_1	V_1	DesvT_1	Inter_2	V_2	DesvT_2	Inter_3	V_3	Interv	Dificultad
109	7	5-7	8		6-8	7				**5-8**	**6,63**

Tabla 42 – Cálculo detallado de la dificultad en el ítem ambiguo en el caso 1 de IRALE.

En la siguiente hoja ItemsIntentados, también del Caso 1, se indica el porcentaje de respuestas correctas obtenidas para cada ítem y el total de valoraciones utilizadas en el cálculo de la dificultad. En el último campo se indica bien el valor obtenido para la dificultad de ese ítem o bien el motivo de eliminación – C.it-1 o C.it-2 – indicando entre paréntesis la iteración en la que se produjo junto con su mensaje aclaratorio.

Item#	IDCallie	PorcResp Correcta	TotalVals	Dificultad/Razón eliminación
1	IRALE1	100	4	1,75
2	IRALE2	100	4	2,25
3	IRALE3	100	4	3,25
4	IRALE4	100	4	1,5
5	IRALE5	100	4	2
6	IRALE6	100	4	5,75
7	IRALE7	100	4	1,75
8	IRALE8	100	4	1,25
9	IRALE9	100	4	5,25
10	IRALE10	100		C.it-2(1): Máximo valoraciones en horquilla (3) menor que el mínimo requerido (4)
11	IRALE11	100	2	3,5
12	IRALE12	100	2	5,5
13	IRALE13	100	4	2,75

Item#	IDCallie	PorcResp Correcta	TotalVals	Dificultad/Razón eliminación
14	IRALE14	100		C.it-2(1): Máximo valoraciones en horquilla (3) menor que el mínimo requerido (4)
15	IRALE15	100	2	2,5
16	IRALE16	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
17	IRALE17	100		C.it-2(1): Máximo valoraciones en horquilla (3) menor que el mínimo requerido (4)
18	IRALE18	100	4	3
19	IRALE19	100	4	4,25
20	IRALE20	100	4	5,5
21	IRALE21	100	4	4,25
22	IRALE22	100	4	5
23	IRALE23	100	6	6,17
24	IRALE24	100		C.it-2(1): Máximo valoraciones en horquilla (4) menor que el mínimo requerido (5)
25	IRALE25	100	8	6,75
26	IRALE26	83,33		C.it-2(1): Máximo valoraciones en horquilla (3) menor que el mínimo requerido (6)
27	IRALE27	100		C.it-2(1): Máximo valoraciones en horquilla (4) menor que el mínimo requerido (6)
28	IRALE28	100	8	7,13
29	IRALE29	100	7	6,71
30	IRALE30	100	8	6,5
31	IRALE31	83,33	6	8
32	IRALE32	100	6	7,83
33	IRALE33	100	8	6,43
34	IRALE34	100	8	6,29
35	IRALE35	100	8	7,29
36	IRALE36	83,33		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
37	IRALE37	100	6	8
38	IRALE38	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
39	IRALE39	100		C.it-2(1): Máximo valoraciones en horquilla (6) menor que el mínimo requerido (8)
40	IRALE40	100	8	6,29
41	IRALE41	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
42	IRALE42	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
43	IRALE43	100	6	8,5
44	IRALE44	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
45	IRALE45	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
46	IRALE46	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
47	IRALE47	100	5	8,6
48	IRALE48	100		C.it-2(1): Máximo valoraciones en horquilla (6) menor que el mínimo requerido (8)
49	IRALE49	83,33	6	8
50	IRALE50	100	6	8,17
51	IRALE51	87,5	8	6,5
52	IRALE52	100		C.it-2(1): Máximo valoraciones en horquilla (6) menor que el mínimo requerido (8)
53	IRALE53	100	8	7,5

Item#	IDCallie	PorcResp Correcta	TotalVals	Dificultad/Razón eliminación
54	IRALE54	66,67		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
55	IRALE55	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
56	IRALE56	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
57	IRALE57	100	6	7,67
58	IRALE58	100	6	8,67
59	IRALE59	100		C.it-2(1): Máximo valoraciones en horquilla (6) menor que el mínimo requerido (8)
60	IRALE60	66,67	6	8,5
61	IRALE61	100		C.it-2(1): Máximo valoraciones en horquilla (3) menor que el mínimo requerido (4)
62	IRALE62	100		C.it-2(1): Máximo valoraciones en horquilla (2) menor que el mínimo requerido (4)
63	IRALE63	100	4	2,5
64	IRALE64	100	4	4,5
65	IRALE65	100	2	4
66	IRALE66	100		C.it-2(1): Máximo valoraciones en horquilla (3) menor que el mínimo requerido (4)
67	IRALE67	100	2	4
68	IRALE68	100	4	4
69	IRALE69	100	4	1,75
70	IRALE70	100	5	5,8
71	IRALE71	100		C.it-2(1): Máximo valoraciones en horquilla (4) menor que el mínimo requerido (6)
72	IRALE72	100		C.it-2(1): Máximo valoraciones en horquilla (4) menor que el mínimo requerido (6)
73	IRALE73	100	6	7,83
74	IRALE74	100	8	7
75	IRALE75	66,67		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
76	IRALE76	100	6	7,33
77	IRALE77	100	6	8,17
78	IRALE78	100	5	8,6
79	IRALE79	100	6	8,33
80	IRALE80	100	6	8,17
81	IRALE81	100		C.it-2(1): Máximo valoraciones en horquilla (4) menor que el mínimo requerido (6)
82	IRALE82	100	4	7,25
83	IRALE83	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
84	IRALE84	100	8	6,86
85	IRALE85	100	6	7,67
86	IRALE86	100	6	7,83
87	IRALE87	100	5	6,2
88	IRALE88	100	8	6,38
89	IRALE89	100	8	6,29
90	IRALE90	60		C.it-2(1): Máximo valoraciones en horquilla (4) menor que el mínimo requerido (5)
91	IRALE91	85,71	7	8,33
92	IRALE92	66,67		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
93	IRALE93	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)

Item#	IDCallie	PorcResp Correcta	TotalVals	Dificultad/Razón eliminación
94	IRALE94	100	4	3,75
95	IRALE95	100		C.it-2(1): Máximo valoraciones en horquilla (2) menor que el mínimo requerido (4)
96	IRALE96	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
97	IRALE97	100		C.it-2(1): Máximo valoraciones en horquilla (3) menor que el mínimo requerido (4)
98	IRALE98	83,33	6	8,5
99	IRALE99	100	7	6,17
100	IRALE100	85,71		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (7)
101	IRALE101	100	6	9,33
102	IRALE102	87,5		C.it-2(1): Máximo valoraciones en horquilla (6) menor que el mínimo requerido (8)
103	IRALE103	83,33	6	8,67
104	IRALE104	100	2	9,5
105	IRALE105	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
106	IRALE106	100	6	7,33
107	IRALE107	83,33		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
108	IRALE108	100	6	8,33
109	IRALE109	87,5	8	6,63
110	IRALE110	100		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (7)
111	IRALE111	75		C.it-2(1): Máximo valoraciones en horquilla (6) menor que el mínimo requerido (8)
112	IRALE112	100	6	9,67
113	IRALE113	75		C.it-2(1): Máximo valoraciones en horquilla (3) menor que el mínimo requerido (4)
114	IRALE114	83,33	6	6
115	IRALE115	100		C.it-2(1): Máximo valoraciones en horquilla (3) menor que el mínimo requerido (4)
116	IRALE116	75		C.it-2(1): Máximo valoraciones en horquilla (3) menor que el mínimo requerido (4)
117	IRALE117	100	4	4,5
118	IRALE118	100	4	4,5
119	IRALE119	100	4	4
120	IRALE120	100		C.it-2(1): Máximo valoraciones en horquilla (4) menor que el mínimo requerido (5)
121	IRALE121	75	4	3,75
122	IRALE122	100		C.it-2(1): Máximo valoraciones en horquilla (3) menor que el mínimo requerido (4)
123	IRALE123	100	2	5
124	IRALE124	100	4	4,5
125	IRALE125	100	2	Responsables calibración IRALE
126	IRALE126	100		C.it-2(1): Máximo valoraciones en horquilla (4) menor que el mínimo requerido (6)
127	IRALE127	75	4	4
128	IRALE128	83,33		C.it-2(1): Máximo valoraciones en horquilla (5) menor que el mínimo requerido (6)
129	IRALE129	60	3	4,33
130	IRALE130	100		C.it-2(1): Máximo valoraciones en horquilla (4) menor que el mínimo requerido (5)
131	IRALE131	100		C.it-2(1): Máximo valoraciones en horquilla (4) menor que el mínimo requerido (5)
132	IRALE132	80		C.it-2(1): Máximo valoraciones en horquilla (4) menor que el mínimo requerido (5)

Tabla 43 – Detalle de los resultados para los 132 ítems en el caso 1 de IRALE.

Las dos últimas tablas reflejan los resultados obtenidos para los 131 ítems no descartados en el Caso 2. En las columnas de los distintos niveles, las celdas resaltadas con el valor de la frecuencia en color negro indican las frecuencias utilizadas por M.dif, mientras que las resaltadas con el valor de la frecuencia en blanco indican las aportaciones descartadas. Los ítems ambiguos aparecen resaltados en esta primera tabla y en la siguiente se detalla su cálculo. La última columna indica la dificultad obtenida para cada ítem en la escala [1,11].

Nivel / Item#	IDCallie	1	2	3	4	5	6	7	8	9	10	11	Dificultad/ Razón
1	IRALE1	2	1	1	0	0	0	0	0	0	0	0	**1,75**
2	IRALE2	1	1	2	0	0	0	0	0	0	0	0	**2,25**
3	IRALE3	0	0	3	1	0	0	0	0	0	0	0	**3,25**
4	IRALE4	3	0	1	0	0	0	0	0	0	0	0	**1,5**
5	IRALE5	1	2	1	0	0	0	0	0	0	0	0	**2**
6	IRALE6	0	0	0	0	2	1	1	0	0	0	0	**5,75**
7	IRALE7	2	1	1	0	0	0	0	0	0	0	0	**1,75**
8	IRALE8	3	1	0	0	0	0	0	0	0	0	0	**1,25**
9	IRALE9	0	0	0	0	3	1	0	0	0	0	0	**5,25**
10	IRALE10	0	0	2	0	1	1	0	0	0	0	0	**3,67**
11	IRALE11	0	0	1	1	0	0	0	0	0	0	0	**3,5**
12	IRALE12	0	0	0	0	1	1	0	0	0	0	0	**5,5**
13	IRALE13	0	1	3	0	0	0	0	0	0	0	0	**2,75**
14	IRALE14	0	0	1	0	2	1	0	0	0	0	0	**4,75**
15	IRALE15	0	1	1	0	0	0	0	0	0	0	0	**2,5**
16	IRALE16	0	0	1	0	2	1	2	0	0	0	0	**6**
17	IRALE17	0	0	1	0	2	0	1	0	0	0	0	**5,67**
18	IRALE18	0	0	4	0	0	0	0	0	0	0	0	**3**
19	IRALE19	0	0	1	1	2	0	0	0	0	0	0	**4,25**
20	IRALE20	0	0	0	0	2	2	0	0	0	0	0	**5,5**
21	IRALE21	0	0	1	1	2	0	0	0	0	0	0	**4,25**
22	IRALE22	0	0	0	0	4	0	0	0	0	0	0	**5**
23	IRALE23	0	0	0	0	1	3	2	0	0	0	0	**6,17**
24	IRALE24	0	0	0	0	1	1	2	0	1	0	0	**6,25**
25	IRALE25	0	0	0	0	1	0	7	0	0	0	0	**6,75**
26	IRALE26	0	0	1	0	1	2	0	0	2	0	0	**5,67**
27	IRALE27	0	0	1	0	0	3	1	0	1	0	0	**6,25**
28	IRALE28	0	0	0	0	0	0	7	1	0	0	0	**7,13**
29	IRALE29	0	0	0	0	1	0	6	0	0	0	0	**6,71**
30	IRALE30	0	0	0	0	2	0	6	0	0	0	0	**6,5**
31	IRALE31	0	0	0	0	0	0	2	2	2	0	0	**8**
32	IRALE32	0	0	0	0	0	0	3	1	2	0	0	**7,83**
33	IRALE33	0	0	1	0	1	2	4	0	0	0	0	**6,43**

Nivel / Item#	IDCallie	1	2	3	4	5	6	7	8	9	10	11	Dificultad/ Razón
34	IRALE34	0	0	0	0	2	1	4	0	1	0	0	6,29
35	IRALE35	0	0	1	0	0	0	5	2	0	0	0	7,29
36	IRALE36	0	0	0	0	0	1	0	4	1	0	0	7,83
37	IRALE37	0	0	0	0	0	0	2	2	2	0	0	8
38	IRALE38	0	0	0	0	0	1	1	2	2	0	0	8,2
39	IRALE39	0	0	0	0	2	1	2	3	0	0	0	7,33
40	IRALE40	0	0	0	0	2	1	4	1	0	0	0	6,29
41	IRALE41	0	0	0	0	0	2	2	1	1	0	0	6,8
42	IRALE42	0	0	0	0	0	1	2	2	1	0	0	7,5
43	IRALE43	0	0	0	0	0	0	1	1	4	0	0	8,5
44	IRALE44	0	0	0	0	0	1	1	0	4	0	0	8,6
45	IRALE45	0	0	0	0	0	1	0	1	4	0	0	8,8
46	IRALE46	0	0	0	0	0	0	1	0	4	1	0	8,83
47	IRALE47	0	0	0	0	0	0	0	2	3	0	0	8,6
48	IRALE48	0	0	0	0	2	0	3	1	2	0	0	7,83
49	IRALE49	0	0	0	0	0	0	2	2	2	0	0	8
50	IRALE50	0	0	0	0	0	0	1	3	2	0	0	8,17
51	IRALE51	0	0	0	0	2	0	6	0	0	0	0	6,5
52	IRALE52	0	0	0	0	2	1	3	1	1	0	0	6,17
53	IRALE53	0	0	0	0	0	0	5	2	1	0	0	7,5
54	IRALE54	0	0	0	0	0	0	1	1	3	0	1	8,4
55	IRALE55	0	0	0	0	0	0	0	2	3	0	1	8,6
56	IRALE56	0	0	0	0	0	1	2	2	1	0	0	7,5
57	IRALE57	0	0	0	0	0	0	3	2	1	0	0	7,67
58	IRALE58	0	0	0	0	0	0	0	2	4	0	0	8,67
59	IRALE59	0	0	0	0	2	0	2	2	2	0	0	8
60	IRALE60	0	0	0	0	0	0	1	1	4	0	0	8,5
61	IRALE61	1	0	1	0	2	0	0	0	0	0	0	4,33
62	IRALE62	1	0	1	0	1	1	0	0	0	0	0	4,67
63	IRALE63	1	0	3	0	0	0	0	0	0	0	0	2,5
64	IRALE64	0	0	1	0	3	0	0	0	0	0	0	4,5
65	IRALE65	0	0	0	2	0	0	0	0	0	0	0	4
66	IRALE66	1	0	1	1	1	0	0	0	0	0	0	4
67	IRALE67	0	0	0	2	0	0	0	0	0	0	0	4
68	IRALE68	0	0	2	0	2	0	0	0	0	0	0	4
69	IRALE69	2	1	1	0	0	0	0	0	0	0	0	1,75
70	IRALE70	0	0	0	0	2	2	1	0	0	0	0	5,8
71	IRALE71	0	0	1	0	1	2	1	1	0	0	0	6,4
72	IRALE72	1	0	1	0	2	2	0	0	0	0	0	5,5
73	IRALE73	0	0	0	0	0	0	3	1	2	0	0	7,83

Nivel Item#	IDCallie	1	2	3	4	5	6	7	8	9	10	11	Dificultad/ Razón
74	IRALE74	0	0	0	1	0	0	7	0	0	0	0	**7**
75	IRALE75	0	0	0	0	1	1	3	1	0	0	0	**6,67**
76	IRALE76	0	0	0	0	0	0	4	2	0	0	0	**7,33**
77	IRALE77	0	0	0	0	0	0	2	1	3	0	0	**8,17**
78	IRALE78	0	0	0	0	0	0	0	2	3	0	0	**8,6**
79	IRALE79	0	0	0	0	0	0	1	2	3	0	0	**8,33**
80	IRALE80	0	0	0	0	0	0	2	1	3	0	0	**8,17**
81	IRALE81	0	0	0	0	0	1	1	1	2	0	1	**8,25**
82	IRALE82	0	0	0	0	0	0	3	1	0	0	0	**7,25**
83	IRALE83	0	0	1	0	1	1	3	0	0	0	0	**6,4**
84	IRALE84	0	0	1	0	0	1	6	0	0	0	0	**6,86**
85	IRALE85	0	0	0	0	0	0	3	2	1	0	0	**7,67**
86	IRALE86	0	0	0	0	0	0	2	3	1	0	0	**7,83**
87	IRALE87	0	0	0	0	1	2	2	0	0	0	0	**6,2**
88	IRALE88	0	0	0	0	2	1	5	0	0	0	0	**6,38**
89	IRALE89	0	0	0	0	2	1	4	1	0	0	0	**6,29**
90	IRALE90	0	0	0	0	0	1	0	0	4	0	0	**9**
91	IRALE91	0	0	0	0	1	0	2	0	4	0	0	**8,33**
92	IRALE92	0	0	0	0	0	0	1	1	3	0	1	**8,4**
93	IRALE93	0	0	0	0	0	1	0	1	4	0	0	**8,8**
94	IRALE94	0	0	2	1	1	0	0	0	0	0	0	**3,75**
95	IRALE95	1	1	0	1	1	0	0	0	0	0	0	**3,67**
96	IRALE96	0	0	0	0	0	0	2	1	2	0	1	**8**
97	IRALE97	1	0	1	0	2	0	0	0	0	0	0	**4,33**
98	IRALE98	0	0	0	0	0	0	1	1	4	0	0	**8,5**
99	IRALE99	0	0	1	0	2	1	3	0	0	0	0	**6,17**
100	IRALE100	0	0	0	0	1	1	3	1	1	0	0	**7,33**
101	IRALE101	0	0	0	0	0	0	0	0	5	0	1	**9,3**
102	IRALE102	0	0	2	0	0	0	3	2	1	0	0	**7,67**
103	IRALE103	0	0	0	0	0	0	0	2	4	0	0	**8,67**
104	IRALE104	0	0	0	0	0	0	0	0	1	1	0	**9,5**
105	IRALE105	0	0	0	0	0	1	1	2	2	0	0	**8,2**
106	IRALE106	0	0	0	0	0	0	4	2	0	0	0	**7,33**
107	IRALE107	0	0	1	0	1	2	2	0	0	0	0	**6,2**
108	IRALE108	0	0	0	0	0	0	1	2	3	0	0	**8,33**
109	IRALE109	0	0	0	0	1	2	4	1	0	0	0	**6,63**
110	IRALE110	0	0	1	0	1	0	2	0	3	0	0	**8,2**
111	IRALE111	0	0	0	0	1	1	2	0	4	0	0	**8,33**
112	IRALE112	0	0	0	0	0	0	0	0	3	2	1	**9,67**
113	IRALE113	0	0	2	0	1	1	0	0	0	0	0	**3,67**

Nivel / Item#	IDCallie	1	2	3	4	5	6	7	8	9	10	11	Dificultad/ Razón
114	IRALE114	0	0	0	0	2	2	2	0	0	0	0	**6**
115	IRALE115	0	0	2	0	1	1	0	0	0	0	0	**3,67**
116	IRALE116	2	0	1	0	1	0	0	0	0	0	0	**1,67**
117	IRALE117	0	0	1	0	3	0	0	0	0	0	0	**4,5**
118	IRALE118	0	0	1	0	3	0	0	0	0	0	0	**4,5**
119	IRALE119	0	0	2	0	2	0	0	0	0	0	0	**4**
120	IRALE120	1	1	2	0	0	1	0	0	0	0	0	**2,25**
121	IRALE121	0	0	2	1	1	0	0	0	0	0	0	**3,75**
122	IRALE122	0	1	1	1	1	0	0	0	0	0	0	**3,5**
123	IRALE123	0	0	0	0	2	0	0	0	0	0	0	**5**
124	IRALE124	0	0	1	0	3	0	0	0	0	0	0	**4,5**
126	IRALE126	0	0	2	0	1	1	2	0	0	0	0	**6,25**
127	IRALE127	0	0	2	0	2	0	0	0	0	0	0	**4**
128	IRALE128	0	0	1	0	1	3	1	0	0	0	0	**6**
129	IRALE129	0	0	1	0	2	0	0	0	0	0	0	**4,33**
130	IRALE130	0	0	2	1	1	1	0	0	0	0	0	**3,75**
131	IRALE131	0	1	3	0	0	0	1	0	0	0	0	**2,75**
132	IRALE132	0	2	2	0	0	0	1	0	0	0	0	**2,5**

Tabla 44 – Frecuencias y dificultades para los ítems del caso 2 de IRALE.

La última tabla refleja el cálculo detallado de la dificultad para los ítems ambiguos del Caso 2.

Item	ValsIni	Inter_1	V_1	DesvT_1	Inter_2	V_2	DesvT_2	Inter_3	V_3	Interv	Dificultad
14	3	3-5	4		4-6	3				**3-6**	**4,75**
17	3	3-5	3	1,0473	5-7	3	1,0277			**5-8**	**5,67**
36	5	6-8	6		7-9	5				**6-9**	**7,83**
42	5	6-8	6		7-9	5				**6-9**	**7,5**
46	5	7-9	6		8-10	5				**7-10**	**8,83**
56	5	6-8	6		7-9	5				**6-9**	**7,5**
62	2	1-3	2		3-5	3		4-6	2	**3-6**	**4,67**
71	4	5-7	5		6-8	4				**5-8**	**6,4**
75	5	5-7	6		6-8	5				**5-8**	**6,67**
95	2	1-3	3	1,2857	2-4	3	1,1157	4-5	2	**2-5**	**3,67**
100	5	5-7	6	1,02	6-8	6	1,0165	7-9	5	**6-9**	**7,33**
109	7	5-7	8		6-8	7				**5-8**	**6,63**
122	3	2-4	4		3-5	3				**2-5**	**3,5**

Tabla 45 – Cálculo detallado de las dificultades en los ítems ambiguos en el caso 2 de IRALE.

Referencias bibliográficas

ADL. (2001). *Scorm overview, advanced distributed learning, version 1.2*: Advanced Distributed Learning.

ADL. (2017a). Sitio Web de ADL SCORM - Sharable Content Object Reference Model. Recuperado 18 / 1 / 2017, desde http://www.adlnet.gov/scorm

ADL. (2017b). Sitio web de Advanced Distributed Learning. Recuperado 18 / 1 / 2017, desde http://www.adlnet.gov/

AENOR. (2010). Norma UNE 71361:2010. Perfil de aplicación LOM-ES para etiquetado normalizado de Objetos Digitales Educativos (ODE). Recuperado 18 / 1 /2017, desde http://www.educaplus.org/documentos/lom-es_v1.pdf

Aguilera, J. M. y González, M. J. (2008). *Test oposiciones cuerpo de tramitación procesal y administrativa de la administración de justicia. Turno libre*: Ed. Cep.

Aguirregoitia, A., Dolado, J. J. y Presedo, C. (2008a). *A landscape metaphor for visualization of software projects*. Presentado en Proceedings of the 4th ACM symposium on Software visualization, Ammersee, Germany.

Aguirregoitia, A., Dolado, J. J. y Presedo, C. (2008b). *A metro map metaphor for visualization of software projects*. Presentado en Proceedings of the 4th ACM symposium on Software visualization, Ammersee, Germany.

Aguirregoitia, A., Dolado, J. J. y Presedo, C. (2009). *Software Visualization Using a Treemap-hybercube Metaphor.* Ponencia presentada en International Conference on DMS - Distributed Multimedia Systems 2009, San Francisco (USA).

Aguirregoitia, A., Dolado, J. J. y Presedo, C. (2010a). *Applying the metro map to software development management.* Ponencia presentada en IS&T/SPIE Electronic Imaging, Visualization and data analysis, San José, CA (USA).

Aguirregoitia, A., Dolado, J. J. y Presedo, C. (2010b). Software Project Visualization Using Task Oriented Metaphors. *Journal of Software Engineering and Applications, JSEA, 3*(11), 1015-1026.

Aguirregoitia, A., Dolado, J. J. y Presedo, C. (2010c). Using the magnet metaphor for multivariate visualization in Software management. *Visual Analytics in Software Engineering - VASE 2009*, 17-24.

Anido, L. y Rodriguez, M. (2002). Estándar para metadatos de objetos educativos. *Learning Technologies Workshop. CEN/ISSS.* Recuperado 7 / 1 / 2016, desde http://www-gist.det.uvigo.es/~lanido/LOMes/LOMv1_0_Spanish.pdf

Arbuckle, J. L. y Wothke, W. (1999). AMOS 4.0 user's guide. Chicago, IL (USA): Smallwaters Corporation.

Armendariz, A. J. (2014). *Métodos de desarrollo dirigidos por modelos y workflows para la calibración psicométrica de ítems: El sistema CALLIE*. (Tesis doctoral), Universidad del País Vasco (UPV/EHU), Bilbao.

Armendariz, A. J., López-Cuadrado, J., Tapias, A., Villamañe, M., Sanz-Lumbier, S. y Sanz-Santamaría, S. (2003). *Learning environments should follow standards: ELSA does.* Ponencia presentada en World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education, Phoenix, Arizona (USA).

Arruabarrena, R. (2005). Filtrado de un banco de ítems (pp. 60). San Sebastián: University of the Basque Country.

Arruabarrena, R. (2010). *E-learning y la calibración de ítems de test: Teoría de Respuesta al Ítem versus calibración basada en juicios de expertos. Un estudio empírico.* (Tesis doctoral), Universidad del País Vasco (UPV/ EHU), San Sebastián.

Arruabarrena, R. y Armendariz, A. J. (2008). Estimación de los parámetros de los ítems de un sistema de e-learning vía expertos (pp. 244). San Sebastián: University of the Basque Country.

Arruabarrena, R. y López-Cuadrado, J. (2006). *Issues to be taken into account when calibrating items.* Ponencia presentada en Current Developments in Technology-Assisted Education Sevilla (España).

Arruabarrena, R., López-Cuadrado, J. y Armendariz, A. J. (2007). Consideraciones para el cómputo de costes de calibraciones de bancos de ítems (pp. 37). San Sebastián: University of the Basque Country.

Arruabarrena, R. y Pérez, T. A. (2005a). Pruebas de campo para calibrar un banco de ítems vía expertos (Informe LSI-TR 03-2005) (pp. 96). San Sebastián (España): Universidad del País Vasco (UPV-EHU).

Arruabarrena, R. y Pérez, T. A. (2005b, 13-16 de septiembre de 2005). *Una experiencia arbitrando incidencias producidas en pruebas de campo.* Ponencia presentada en VI congreso nacional de Informática Educativa. I Simposio Nacional de Tecnologías de la Información y las Comunicaciones en la Educación: Sintice-2005 (ADIE, CEDI'05), Granada.

Arruabarrena, R. y Pérez, T. A. (2010). Calibración de ítems con expertos: procesos BPM, ejecución, análisis y mejora. Una investigación empírica (pp. 53). San Sebastián: University of the Basque Country.

Arruabarrena, R., Pérez, T. A., Gutiérrez, J., López-Cuadrado, J. y Vadillo, J. A. (2002). *On Evaluating Adaptive Systems for Education.* Ponencia presentada en AH2002, 2nd. International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Málaga.

Arruabarrena, R., Sanz-Santamaría, S. y Gutiérrez, J. (2007, 11-14 sep.t). *Desarrollo eficiente de pruebas de campo.* Ponencia presentada en 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).

Arruabarrena, R., Vadillo, J. A. y Gutiérrez, J. (2003). *Are Experts Difficulty Guessing And Statistical Results Comparable?* Ponencia presentada en Advances in technology-based education: towards a knowledge-based society Badajoz (España).

ASC. (1996). RASCAL - Rasch Analysis Program. St. Paul, Minnesota (USA): Assessment Systems Corporation.

ASC. (2015). XCALIBRE 4 - Software for IRT analysis and calibration. St. Paul, Minnesota (USA): Assessment Systems Corporation.

ASPECT. (2017). Sitio web de ASPECT - Adopting Standards and Specifications for Educational content. Recuperado 18 / 1 / 2017, desde http://www.aspect-project.org/node/71

Baker, F. B. (1992). *Item response theory: parameter estimation techniques*. New York (USA): Marcel Dekker.

Baker, F. B. (2001). *The basics of item response theory*. University of Maryland, College Park (USA): ERIC Clearinghouse on Assessment and Evaluation.

Barbero, M. I. (1996). Banco de ítems. En J. Muñiz (Ed.), *Psicometría* (pp. 139-170). Madrid (España): Editorial Universitas, S.A.

Barbero, M. I. y Navas, M. J. (1995). Creación de un sistema computerizado de evaluación de la capacidad matemática. Madrid (Spain): Centro de Investigación, Documentación y Evaluación (CIDE).

Bentler, P. M. (1985). Theory and implementation of EQS: a structural equations program. Los Angeles, CA (USA): BMDP Statistical Software.

Bentler, P. M. y Wu, E. J. C. (1993). *EQS user's guide*. Los Angeles, CA (USA): BMDP Statistical Software.

Binet, A. y Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année psychologique, 11*, 191-244.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F. M. Lord y M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. chapters 17-20). Reading (USA): Addison-Wesley.

Bock, R. D. y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters. An application of an EM algorithm. *Psychometrika, 35*, 179-197.

Brown, K.-K. y Brown, J. D. (2000). The Japanese Placement Tests at the University of Hawaii: applying item response theory (pp. 89). Honolulu, Hawai (USA): Second Language Teaching & Curriculum Center (University of Hawaii).

Büchner, A. (2011). *Moodle 2 Administration*. Birmingham (UK): Ed. Packt Publishing Ltd.

Bunderson, C. V., Inouye, D. K. y Olsen, J. B. (1989). The four generations of computerized educational measurement. En R. L. Linn (Ed.), *Educational Measurement*. London (UK): Collier Macmillan Publishers.

Burke, N. W., Kaufman, B. D. y Webb, N. L. (1985). The Wisconsin item bank: development, operation and related issues: Madison Wisconsin Department of Public Instruction.

Caro, J. L. (1988). *Eficacia de las pruebas objetivas para la enseñanza de las técnicas de expresión gráfica en la ingeniería.* UPV-EHU, Bilbao.

Case, R., Demetriou, A., Platsidou, M. y Kazi, S. (2001). Integrating concepts and tests of intelligence from the differential and developmental traditions. *Intelligence, 29*(4), 307-336.

Clarenc, C. A. (2012). *Tipos de LMS: Características Requisitos - Procedimientos para seleccionar un LMS.* Recuperado 18 / 1 / 2017 desde http://es.scribd.com/doc/100084611/Tipos-de-LMS-Caracteristicas-Requisitos-Procedimientos-para-seleccionar-un-LMS

Clarenc, C. A. (2013). *Trabajo y aprendizaje colaborativos: Mejores prácticas y estrategias*. Congreso Virtual Mundial de e-learning. Recuperado 18 / 1 / 2017, desde http://es.scribd.com/doc/189219329/Trabajo-y-aprendizaje-colaborativos-Mejores-practicas-y-estrategias

Clarenc, C. A., Castro, S. M., López de Lenz, C., Moreno, M. E. y Tosco, N. B. (2013). *Analizamos 19 plataformas de e-Learning: Investigación colaborativa sobre LMS*. Grupo GEIPITE, Congreso Virtual Mundial de e-Learning. Recuperado 18 / 1 / 2017, desde www.congresoelearning.org

Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-De-La-Cruz, J. L. y Ríos, A. (2004). SIETTE: A Web-Based Tool for Adaptive Testing. *International Journal of Artificial Intelligence in Education, 14*, 1-33.

Conejo, R., Guzmán, E. y Pérez, J. L. (2008). *Un Estudio sobre la Dificultad de los Ítems en Tests de Informática.* Ponencia presentada en XIV Jornadas de Enseñanza Universitaria de la Informatica Granada (España).

Cornish, G. y Wines, R. (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: Cambridge University Press.

Chang, S.-W., Hanson, B. A. y Harris, D. J. (2000). *A standardization approach to adjusting pretest item statistics.* Ponencia presentada en Annual meeting of the National Council on Measurement in Education, New Orleans (USA).

Choppin, B. H. (1981). Educational measurement and the item bank model. En C. Lacey y D. Lawton (Eds.), *Issues in evaluation and accountability*. London: Methuen & Co.

Dalkey, N. C., Brown, B. y Cochran, S. (1970). The Delphi method, III. Use of self rating to improve group estimates. *Technological Forecasting and Social Change, 1*, 283-291.

DC-SingaporeFR. (2017). Sitio Web de Singapore Framework for Dublin Core Application Profiles. Recuperado 18 / 1 / 2017, desde http://dublincore.org/documents/singapore-framework

DCMI. (2017). Sitio Web de Dublin Core Metadata Initiative. Recuperado 18 / 1 / 2017, desde http://dublincore.org/

Dix, A., Finlay, J., Abowd, G. y Beale, R. (1998). *Human-Computer Interaction, 2nd edition*: Pearson Education Limited.

Douglas, J. P. (1980). *Applying laten trait theory to a classroom examination system: Model comparison and selection.* Presentado en Comunicación presentada en la reunión anual de la AERA (American Educational Research Association), Boston.

Elliot, C. D. (1983). British ability scales. Manuals 1-4. Windsor, England: NFER-Nelson.

Emenogu, B. C. y Childs, R. A. (2005). Curriculum, translation, and differential functioning of measurement and geometry items. *Canadian Journal of Education, 28*(1&2), 128-146.

Fallon, C. y Brown, S. (2003). *E-learning Standards: A guide to purchasing, developing and deploying standards-conformant e-learning*: Ed. ST. Lucie Press.

Fernández-Manjón, B., Moreno-Ger, P., Sierra, J. L. y Martínez-Ortiz, I. (2007). Uso de estándares aplicados a TIC en Educación. Informe Nº 16. *CNICE*.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: John Wiley.

Fleiss, J. L. y Cohen, J. (1973). The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement, 33*, 613-619.

Frechtling, J. A. y Sharp, L. M. (1997). *User-friendly handbook for mixed method evaluations*: DIANE Publishing.

García-Viñas, J. I., Conejo, R., Gastón, A., López, C. y Ropero, C. (2013). *La plataforma siette una herramienta para el aprendizaje de la botánica forestal*. Presentado en 6º Congreso forestal español, Vitoria.

Gutiérrez, J., Pérez, T. A., Usandizaga, I. y Lopistéguy, P. (1996). *HyperTutor: Adapting Hypermedia Systems to the User*. Ponencia presentada en International Conference on User Modeling, UM, Kailua-Kona, USA.

HABE. (1984). Helduen euskalduntzearen oinarrizko kurrikulua (HEOK). San Sebastián (España): Eusko Jaurlaritza - Gobierno Vasco.

Haley, D. C. (1952). Estimation of the dosage mortality relationship when the dose is subject to error (Technical Report No. 15). Stanford (USA): Applied Mathematics and Statistics Laboratory, Stanford University.

Hambleton, R. K. y Swaminathan, H. (1985). *Item response theory : principles and applications*. Boston (USA): Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H. y Rogers, H. J. (1991a). *Fundamentals of Item Response Theory* (Vol. 2). Newbury Park, CA: Sage.

Hambleton, R. K., Zaal, J. N. y Pieters, J. P. M. (1991b). Computerized adaptive testing: theory, applications, and standards. En R. K. Hambleton y J. N. Zaal (Eds.), *Advances in educational and psychological testing*. Norwell, Massachussets (USA): Kluwer Academic Publishers.

Hankins, J. A. (1990). Florida's vocational testing program. *Journal of Employment Counseling, 27*(3), 130-138.

Harvey, J. (1998). *Evaluation Cookbook* (Jen Harvey (maintained by Phil Barker) ed.). Heriot Watt University Edinburgh: Learning Technology Dissemination Initiative, Institute for Computer Based Learning.

Heard, J. E., Byrne, D. E. y Ward, J. P. T. (2002). Computerised adaptive testing: results of a trial at King's College London. *Journal of Physiology, 544*, 313.

Helmer, O. y Rescher, N. (1959). On the epistemology of the inexact sciences. *Management Science, 6*, 25-52.

Hendrickson, A. B. y Kolen, M. J. (1999). IRT equating of the MCAT (pp. 23). Washington, DC (USA): Association of American Medical Colleges.

Henning, G. (1986). Item banking via dBase II: The UCLA ESL Proficiency Examination experience. En C. W. Stansfield (Ed.), *Technology and language testing* (pp. 69-77). Washington, DC (EE:UU): TESOL (Teachers of English to Speakers of Other Languages, Inc.).

Hernández, E. (2003). Estándares y especificaciones *e-learning*: ordenando el desorden. Recuperado 18 / 1 / 2017, desde http://www.uv.es/ticape/docs/eduardo.pdf

Hill, P. W. (1985). *The tests of reading comprehension (TORCH).* Ponencia presentada en la reunión anual de la IEA Oxford (Gran Bretaña).

Hiscox, M. D. y Brzenzinski, E. (1980). *A guide to item banking in education (prefared for the Annual conference on large-scale assessment)*. Portland, OR:.

Ho, R.-G., Wang, H.-Y. y Shyu, H.-J. (1997). *An adaptive strategy on keyboarding: using adaptive drill system model.* Ponencia presentada en 6th International Conference on Computer Assisted Instruction, Taipei (Taiwan).

Hontangas, P., Olea, J., Ponsoda, V., Revuelta, J. y Wise, S. L. (2004). Assisted self-adapted testing: a comparative study. *European Journal of Psychological Assessment, 20*(1), 2-9.

Horton, W. y Horton, K. (2003). *E-Learning tools and technologies: A consumer's guide for trainers, teachers, educators, and instructional designers*. Indianapolis, IN: Wiley publishing.

Huang, C. H., Kalohn, J. C., Lin, C. J. y Spray, J. A. (2000). Estimating item parameters from classical indices for item pool development with a computerized classification test (ACT Research Report Series 2000-4). Iowa City, Iowa (USA).

IEEE. (2002). Draft standard for Learning Object Metadata P1484.12.1-2002. Recuperado 18 / 1 / 2017, desde biblio.educa.ch/sites/default/files/20130328/lom_1484_12_1_v1_final_draft_0.pdf

IEEE. (2017). Sitio Web del IEEE - Institute of Electrical and Electronical Engineers. Recuperado 18 / 1 / 2017, desde http://www.ieee.org/index.html

IMS-CC. (2017). Sitio web del IMS GLC: Common Cartridge. Recuperado 18 / 1 / 2017, desde https://www.imsglobal.org/activity/common-cartridge

IMS-CP. (2001). *IMS Content Packaging Best Practice Guide. Version 1.1.2 Final Specification.* Recuperado 18 / 1 / 2017 desde http://www.imsproject.org/content/packaging/cpv1p1p2/imscp_bestv1p1p2.html

IMS-CP. (2017). Sitio web de IMS GLC: Content Packaging Specification. Recuperado 18 / 1 / 2017, desde http://www.imsglobal.org/content/packaging/

IMS-GLC. (2017). Sitio Web del IMS GLC: Instructional Management Systems Global Learning Consortium. Recuperado 18 / 1 / 2017, desde http://www.imsglobal.org/

IMS-QTI. (2002). *IMS Question and Test Interoperability: Version 1.2 - Final Specification.* I. IMS Global Learning Consortium (Ed.) Recuperado 18 / 1 / 2017 desde http://www.imsglobal.org/question/index.html#version1.2

IMS-QTI. (2017). Sitio web de IMS GLC: Global Question & Test Interoperability (QTI) Specification Recuperado 18 / 1 / 2017, desde http://www.imsglobal.org/question/

Irastorza, J. (2014). *CALLIE-EXT.* E.U. Ingeniería Técnica Industrial, Bilbao.

Jeffries, R., Miller, J. R., Wharton, C. y Uyeda, K. M. (1991). *User interfaces evaluation in the real world: A comparison of four techniques.* Ponencia presentada en ACM Humman Computer Interaction Conf.

Karat, C. M., Campbell, R. y Fiegel, T. (1992). *Comparison of empirical testing and walkthrough methods in user interface evaluation.* Ponencia presentada en Human Factors in Computing Systems Conf., New York.

Kingsbury, G. G. y Weiss, D. J. (1983). A comparison of IRT-based Adaptive Mastery Testing and Sequential Mastery Testing Procedure. En D. J. Weiss (Ed.), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing* (pp. 257-283). New York: Academic Press.

Kolen, M. J. y Brennan, R. L. (1995). *Test equating: methods and practices*. New York (USA): Springer-Verlag.

Kubeš, T. (2007). *Application of Hypermedia Systems in e-Learning.* Czech Technical University in Prague., Praga (R. Chequia). Recuperado de http://www.tomaskubes.net/eat/

Kubinger, K. D. (1985). *On a Rash model based test for noncomputerized adaptative testing.* Ponencia presentada en 13th IPN Conference on Latent Trait and Latent Class Models in Educational Research, Kiel.

Landis, J. R. y Koch, G. G. (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 foundations of latent structure analysis*. Princenton: Princenton University Press.

Leyva, J. (2010). Listado de comparativas, informes y evaluaciones entre Moodle, Sakai, Blackboard y otros LMS. Recuperado 18 / 1 / 2017, desde http://openlearningtech.blogspot.com.es/2010/02/listado-de-informes-evaluaciones-y.html

Li, Y. H. y Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement, 29*(1), 3-25.

Lin, L.-C., Tseng, H.-M. y Wu, S.-C. (1999). Item analysis of the registered nurse licensure exam taken by nurse candidates from vocational nursing high schools in Taiwan. *Prooceedings of the National Science Council, Republic of China (Part D), 9*(1), 24-31.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. y Wright, B. D. (1997). *A user's guide to BIGSTEPS. Rasch model computer program*. Chicago, IL (USA): MESA Press.

Lofgren, R. E. (2005). Validity evidence of a multiple-choice test and a performance test in an employment setting (pp. 164). Pittsburgh, PA (USA): University of Pittsburgh.

López-Cuadrado, J. (2008). *Evaluación mediante test adaptativos informatizados en el contexto de un sistema adaptativo para el aprendizaje de la lengua vasca*. Univ. País Vasco/ Euskal Herriko Unibertsitatea, San Sebastián.

López-Cuadrado, J. (2010). *Test adaptativos informatizados de ingreso en un sistema e-learning*. Saarbrüken: LAP Lambert Academic Publishing.

López-Cuadrado, J. y Armendariz, A. J. (2006). Obtención de estimaciones de los parámetros durante la calibración de un banco de ítems (Informe LSI-TR 13-2007) (pp. 271). San Sebastián (España): Lenguajes y Sistemas Informáticos (Universidad del País Vasco UPV-EHU).

López-Cuadrado, J., Armendariz, A. J., Latapy, M. y Lopistéguy, P. (2008). A genre-based perspective for the development of communicative computerized adaptive tests. *Journal of Educational Technology & Society, 11*(1), 87-101.

López-Cuadrado, J., Armendariz, A. J. y Pérez, T. A. (2003). ADISTI: an authoring tool for creating and managing exercises in e-learning systems. En A. Méndez Vilas, J. A. Mesa González y J. Mesa González (Eds.), *Advances in technology-based education: towards a knowledge-based society* (Vol. 3, pp. 1555-1559). Badajoz (España): Junta de Extremadura (CECT).

López-Cuadrado, J., Armendariz, A. J. y Pérez, T. A. (2006). *Adaptive evaluation in an e-learning system architecture.* Ponencia presentada en Current Developments in Technology-Assisted Education Sevilla (España).

López-Cuadrado, J., Armendariz, A. J., Pérez, T. A. y Arruabarrena, R. (2008). *Helping tools for item bank calibration and development of computerized adaptive tests.* Ponencia presentada en International Technology, Education, and Development Conference (INTED2008), Valencia (España).

López-Cuadrado, J., Armendariz, A. J., Pérez, T. A., Arruabarrena, R. y Vadillo, J. A. (2009). Computerized adaptive testing, the item bank calibration and a tool for easing the process *Technology Education and Development* (pp. 457-478): Eds. A. Lazinica & C. Calafate.

López-Cuadrado, J., Armendariz, A. J., Presedo, C., Segundo, U., Barrena, R., Korta, J. (2015). *Why tables on clinical practice guidelines are not easily computerizable.* Ponencia presentada en HTAi 12th Annual Meeting, 15-17 June, 2015, Oslo (Norway).

López-Cuadrado, J., Pérez, T. A. y Armendariz, A. J. (2005). Evaluación mediante test: ¿Por qué no usar el ordenador? *Revista Iberoamericana de Educación, 36*(11), 1-15.

López-Cuadrado, J., Pérez, T. A., Sanz-Santamaría, S., Armendariz, A. J., Gutiérrez, J. y Vadillo, J. Á. (2007). *Improvement, Standardization, Extensibility, Helping Tools: Key Issues to Build a Successful E-Learning System (Some Innovations for Keeping an e-Learning System Alive).* Ponencia presentada en International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CIS2E 07), Virtual Conference.

López-Cuadrado, J., Pérez, T. A., Vadillo, J. A. y Arruabarrena, R. (2002). *Integrating Adaptive Testing in an Educational System.* Ponencia presentada en Educational Technology in Cultural Context: ETCC2002. First International Conference on, Joensuu, Finland.

López-Cuadrado, J., Pérez, T. A., Vadillo, J. A. y Gutiérrez, J. (2010). Calibration of an item bank for the assessment of Basque language knowledge. *Computers & education, 55*(3), 1044-1055.

López Alonso, C. y Matesanz del Barrio, M. (2009). *Las plataformas de aprendizaje: del mito a la realidad*. Madrid: Biblioteca Nueva.

López Pina, J. A. (1995). *Teoría de la respuesta al ítem: fundamentos*. Murcia (España): DM-PPU.

Lord, F. M. (1968). An analysis of the verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1020.

Lord, F. M. (1974). Individualized testing and item characteristics curve theory. En D. H. Krantz, R. C. Atkinson, R. D. Luce y P. Suppes (Eds.), *Contemporary developments in mathematical psychology, Vol. II*. San Francisco, California (USA): Freeman.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, N.J.: L. Erlbaum Associates.

Lord, F. M. y Novick, M. (1968). *Statistical Theories of Mental Tests Scores*. New York: Addison Wesley.

Lynch, D. F. (2004). Linking strategy, structure, process, and performance in integrated logistics. *Journal of Business Logistics, 25*(2), 65-94.

Marcelo, C. (2006). *Prácticas de e-learning*. Sevilla: Octaedro Andalucía.

Marchesi, Á. (2001). Redes de escuelas para la evaluación y el cambio educativo. Manchester (UK): The Office for Standards in Education.

Mark, M. A. y Greer, J. E. (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 Moreno-Rodríguez, R. (2002). *Construcción de un banco de ítems informatizado para la evaluación de conocimientos sobre una materia universitaria.* Ponencia presentada en Proc. III Jornadas Andaluzas de Calidad en la Enseñanza Universitaria, Universidad de Sevilla.

MEC. (2007). PIRLS 2006. Informe español. *Secretaría General Técnica* (pp. 107). Madrid: Catálogo oficial de publicaciones oficiales.

Millman, J. y Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement*(21), 315-330.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195.

Mislevy, R. J. y Bock, R. D. (1984). *BILOG I maximum likelihood item analysis and test scoring: logistic model*. Mooresville, IN (USA): Scientific Software International.

Mislevy, R. J. y Bock, R. D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models*. Mooresville, IN (USA): Scientific Software International.

Mislevy, R. J. y Bock, R. D. (1990). *BILOG 3*. Mooresville, IN (USA): Scientific Software International.

Mislevy, R. J., Bock, R. D. y Muraki, E. (1988). *BIMAIN*. Mooresville, IN (USA): Scientific Software International.

Moodle. (2017). Sitio Web de Moodle (foros, documentación). Recuperado 18 / 1 / 2017, desde http://moodle.org/

Moore, W. P. (1994). The Devaluation of Standardized Testing: One District's Response to a Mandated Assessment. *Journal of Applied Measurement in Education, 7*(4), 343-367.

Mostow, J., Tobin, B. y Cuneo, A. (2002). *Automated comprehension assessment in a reading tutor.* Ponencia presentada en Evidence-centered design (ECD) to approach to creating diagnostic e-assessments (6th International Conference on Intelligent Tutoring Systems, ITS2002 Workshop), San Sebastián (España).

Mullis, I. V. S., Martin, M. O., Kennedy, A. M. y Foy, P. (2007). *PIRLS 2006. International Report*. Boston College, MA 02467 (United States): TIMSS & PIRLS Intenational Study Center - IEA.

Muñiz, J. (1996). *Psicometría*. Madrid (España): Editorial Universitas, S.A.

Muñiz, J. (2000). *Teoría clásica de los Tests*. Madrid: Ediciones Pirámide.

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems. *Papeles del Psicólogo, 31*(1), 57-66.

Murphy, R. (2002). *A review of South African research in the field of dynamic assessment*. Pretoria (South Africa): University of Pretoria.

Narayan, D. y Cassidy, M. F. (2001). A dimensional approach to measuring social capital: development and validation of a social capital inventory. *Current Sociology, 49*(2), 59-102.

Nielsen, J. (1993). *Usability Engineering* (Vol. Biblioteca): AP Professional (Academic Press).

Nielsen, J. y Mack, R. L. (1994). *Usability Inspection Methods*. New York: John Wiley &Sons, Inc.

Nitko, A. J. y Hsu, T. C. (1984). A comprehensive microcomputer system for classroom testing. *Journal of Educational Measurement, 21*, 377-390.

NJDoE. (2006). 2004 grade eight proficiency assessment (GEPA) technical report (pp. 142). New Jersey (USA): New Jersey Department of Education.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*, 1-18.

Nunnally, J. C. (1978). *Psychometric theory*. New York (USA): McGraw-Hill.

O'Brien, M. L. y Hampilos, J. O. (1988). The feasibility of creating an item bank from a teacher-made test using the Rash model. *Educational and Psychological Measurement*(48), 201-212.

OECD. (2012). *PISA 2009 Technical Report*: OECD Publishing.

OECD. (2013). *PISA 2012 Assessment and Analytical Framework*: OECD Publishing.

OECD. (2014). *PISA 2012 Technical Report*: OECD Publishing.

Olea, J., Abad, F. J., Ponsoda, V. y Ximénez, M. C. (2004). Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: diseño y comprobaciones psicométricas. *Psicothema, 16*(3), 519-525.

Olea, J. y Ponsoda, V. (2003). *Tests adaptativos informatizados*. Madrid (España): Ediciones UNED.

Otero, M. C. (2003). *Evaluación empírica de la comprensión del modelado dinámico en los lenguajes UML y OML de aplicaciones software.* UPV/EHU.

Otero, M. C. y Dolado, J. (2007). Conocimiento empírico a través de una familia de experimentos. En E. J. Tuya, I. Ramos y J. Dolado (Eds.), *Técnicas cuantitativas para la gestión en la ingeniería del software* (pp. 155-180). Oleiros-A Coruña (España): NETBIBLO.

Ozen, D. J. y Reise, S. P. (1994). *Personality assessment*. Palo Alto.

Palacios, F., Pérez, E., Callejón, J. y Herrerías, R. (1999). *Un método para contrastar la bondad de un experto en la metodología PERT*. Ponencia presentada en XIII Reunión Anual de ASEPELT.

Pérez, C. (1999). *Técnicas de muestreo estadístico. Teoría, práctica y aplicaciones informáticas*. Madrid (Spain): ra-ma.

Pérez, T. A. (2000). *Un hiperentorno adaptativo para el aprendizaje instructivo / constructivo.* Univ. País Vasco / Euskal Herriko Unibertsitatea, San Sebastián.

PLS-RAMBOLL. (2004). Studies in the Context of the E-learning Initiative: Virtual Models of European Universities. Recuperado 18/ 1 / 2017, desde http://tecnologiaedu.us.es/cuestionario/bibliovir/pls.pdf

Pollit, A. B. (1985). Item banking and school measurement. En W. Entwistle (Ed.), *New Directions in Educational Psychology: Learning and teaching*. East Sussex (England): The Falmer Press.

Presedo, C., Armendariz, A. J. y López-Cuadrado, J. (2012). *Calibración de ítems para test informatizados: Descripción detallada de las fases en la construcción de test de evaluación adaptativos mediante ordenador*: Editorial Académica Española.

Presedo, C. y Dolado, J. J. (2006). *El problema de la coordinación en proyectos software: Enfoque mediante sistemas Multiagente.* Ponencia presentada en XI Jornadas de Ingeniería del Software y Bases de Datos, Sitges (España).

Presedo, C. y Dolado, J. J. (2007). Medición práctica de la coordinación utilizando GQ(I)M y CMMi. *Actas del 8º Taller sobre Apoyo a la Decisión en Ingeniería del Software*.

Presedo, C., Dolado, J. J. y Aguirregoitia, A. (2010). Estudio de métricas para el control de proyectos software. *Actas del 10° Taller de las Jornadas sobre apoyo a la decisión en Ingeniería del Software y Bases de Datos, 4*(1), 65-72.

Prieto, G. y Delgado, A. R. (1999). Medición cognitiva de las aptitudes. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests informatizados: fundamentos y aplicaciones* (pp. 207-226). Madrid (España): Ediciones Pirámide.

Rabeneck, L., Cook, K. F., Wristers, K., Souchek, J., Menke, T. y Wray, N. P. (2001). SODA (severity of dyspepsia assessment): a new effective outcome measure for dyspepsia-related health. *Journal of Clinical Epidemiology, 54*, 755-765.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.

Ree, M. J. (1977). Implementation of a model adaptive testing system at an Armed Forces Entrance and Examining Station. En D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference (NTIS No AD-A060 049)* (pp. 216-220). Minneapolis, Minnesota (USA): Psychometrics Methods Program, Department of Psychology (University of Minnesota).

Renom, J. (1992). *METRIX: Manual del usuario*. Barcelona (España): Engine.

Renom, J. y Doval, E. (1999). Tests adaptativos informatizados: estructura y desarrollo. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests informatizados: fundamentos y aplicaciones* (pp. 127-162). Madrid (España): Ediciones Pirámide.

Revuelta, J. y Ponsoda, V. (1998). Un test adaptativo informatizado de análisis lógico basado en la generación automática de ítems. *Psicothema, 10*(3), 709-716.

Riley, K. y Mills, D. (2008). *IMS Common Cartridge (CC) Authorization Web Service v1.0 Final Specification* I. IMS Global Learning Consortium (Ed.) Recuperado 18 / 1 / 2017 desde www.imsglobal.org/cc/

Rios, A., Perez-de-la-Cruz, J. y Conejo, R. (1998). *SIETTE: Intelligent Evaluation System using Test for TeleEducation.* . Ponencia presentada en 4th International Conference on Intelligent Tutoring System. ITS'98. Workshops papers., San Antonio,Texas (USA).

Rojas Tejada, A. J. (2001). Pasado, presente y futuro de los tests adaptativos informatizados: entrevista con Isaac I. Béjar. *Psicothema, 13*(4), 685-690.

Roncel Vega, V. M. (2007). Aprende-le: inventario de estrategias de aprendizaje para la lengua española. *redELE: red electrónica de didáctica del Español como Lengua Extranjera, 9*, 14.

Scanlan, D. A. (1989, september). Structured Flowcharts Outperform Pseudocode: An Experimental Comparison. *IEEE Software, 6*, 28-36.

Scriven, M. (1991). *Evaluation Thesaurus. 4th edition*: Sage Publications.

Shen, L. (1993). *Constructing a measure for longitudinal medical achievement studies by the Rasch model one-step equating.* Ponencia presentada en Annual meeting of the American Educational Research Association, Atlanta, GA (USA).

Shneiderman, B. (1998). *Designing the User Interface, 3rd edition*. Reading-Massachusetts: Addison Wesley Longmang, Inc.

Singh, J. (2004). Tackling measurement problems with Item Response Theory principles, characteristics, and assessment, with an illustrative example. *Journal of Business Research, 57*(2), 184-208.

Sjøberg, S. (2004, 02-04-2004). *Science Education: The voice of the learners.* Ponencia presentada en Proc. Conference on Increasing Human Resources for Science and Technology in Europe, Bruselas (UE).

Smythe, C. y Nielsen, B. (2007). *IMS Content Packaging (CP) Specification Primer v1.2 Public Draft.* I. IMS Global Learning Consortium (Ed.) Recuperado 18 / 1 / 2017 desde www.imsglobal.org/content/packaging/

Smythe, C., Shepherd, E., Brewer, L. y Lay, S. (2002). *IMS Question & Test Interoperability: ASI Information Model Specification. Final Specification Version 1.2.* I. IMS Global Learning Consortium (Ed.) Recuperado 18 / 1 / 2017 desde http://www.imsglobal.org/question/qtiv1p2/imsqti_asi_infov1p2.html

Sommer, M., Arendasy, M. y Häusler, J. (2005). *Theory-based construction and validation of a modern computerized intelligence test battery*. Ponencia presentada en International Military Testing Association 47th Annual Conference, Singapore (Republic of Singapore).

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5*, 417-426.

Stage, C. (2003). Teoría clásica de medición o teoría de respuesta al ítem. La experiencia sueca. *eJournal Estudios Públicos, nº 90*, 185-217.

Stocking, M. L. (1994). Three practical issues for modern adaptive testing item pools (Report No. ETS-RR-94-5). Princeton, New Jersey (USA): Educational Testing Service.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D. y Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: effects of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement, 27*(1), 27-51.

Tapias, J. A. (2008). *1.112 preguntas básicas para oposiciones a bombero*: Ed. Cep.

Tessmer, M. (1993). *Planning and Conducting: Formative Evaluations*. London: Kogan Page Limited.

Thissen, D. M. y Mislevy, R. J. (2000). Testing algorithms. En H. Wainer (Ed.), *Computerized adaptive testing: a primer (second edition)* (pp. 101-132). Mahwah, New Jersey (USA): Lawrence Erlbaum Associates.

Thompson, N. A. y Guyer, R. (2010). *User´s Manual for ITEMAN 4.1. Classical Item and test analysis.* St. Paul, Minnesota (USA): Assessment Systems Corporation.

Thompson, N. A. y Weiss, D. J. (2006). Item response theory parametrization of the Multistate Bar Exam (pp. 41). Minnesota (USA): University of Minnesota.

Tognolini, J. (1982). Pupil achievement in stage 6 mathematics. (Discussion paper N. 15): Perth: Education Department of Western Australia.

Trella, M., Carmona, C. y Conejo, R. (2005). *MEDEA: an open service-based learning platform for developing intelligent educational systems for the web.* Ponencia presentada en Workshop on Adaptive Systems for Web-Based Education: Tools and Reusability (AIED'05), Amsterdam, the Netherlands.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1-13.

Umar, J. (1997). Item banking. En J. Keeves (Ed.), *Educational research, methodology, and measurement (2nd edition)* (pp. 923-930). Oxford: Pergamon.

Urry, V. W. (1978). *ANCILLES: item parameter estimation program with normal ogive and logistic three-parameter model options*. Washington, DC (USA): Civil Service Commission, Development Center.

Vale, C. D. y Gialluca, K. A. (1985). *ASCAL: A microcomputer program for estimating logistic IRT item parameters (RR ONR 85-4)*. St. Paul, Minnesota (USA): Assessment Systems Corporation.

van der Linden, W. J. (1986). Test item banking. *Applied Psychological Measurement, 10*(4), 332-343.

van Thiel, C. C. y Zwarts, M. A. (1986). Development of a Testing Service System *Journal of Applied Psychological Measurement*, *10*(4), 391-403.

Vispoel, W. P. (1999). Creating computerized adaptive tests of music aptitude: problems, solutions, and future directions. En F. Drasgow y J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 151-176). Mahwah, New Jersey (USA): Lawrence Erlbaum Associates.

Wainer, H., Bradlow, E. T. y Du, Z. (2000). Testlet response theory: an analog for the 3PL model useful in testlet-based adaptive testing. En W. J. van der Linden y C. A. W. Glas (Eds.), *Computerized adaptive testing: theory and practice* (pp. 245-269). Dordrecht (The Netherlands): Kluwer Academic Publishers.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L. (2000). Future challenges *Computerized Adaptive Testing: A Primer (2nd*

edition) (1st ed., pp. 231-269). Hillsdale, New Jersey: Lawrence Erlabum Associates.

Wainer, H. y Mislevy, R. J. (1990). Item response theory, item calibration and proficiency estimation. En H. Wainer (Ed.), *Computerized adaptive testing: a primer* (pp. 65-102). Hillsdale, New Jersey (USA): Lawrence Erlbaum Associates.

Wainer, H. y Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. En H. Wainer (Ed.), *Computerized adaptive testing: a primer (second edition)* (pp. 61-99). Mahwah, New Jersey (USA): Lawrence Erlbaum Associates.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17-26.

Weber, G., Kuhl, H.-C. y Weibelzahl, S. (2001). Developing adaptive internet based course with the authoring system NetCoach. En S. Reich, M. Tzagarakis y P. De Bra (Eds.), *LNCS. Revised Papers from the nternational Workshops OHS-7, SC-3, and AH-3 on Hypermedia: Openness, Structural Awareness, and Adaptivity* (Vol. 2266, pp. 226-238). London (UK): Springer-Verlag

Weibelzahl, S. (2002). *Evaluation of Adaptive Systems. (Ph dissertation).* University of Trier, Freiburg (Alemania).

Weiss, D. J. y Yoes, M. E. (1991). Item response theory En R. K. Hambleton y J. N. Zaal (Eds.), *Advances in Educational and Psychological Testing: Theory and Applications* (pp. 69-95). Norwell (Netherlands): Kluwe Academic.

William, H. (2008). *Moodle 1.9 E-learning course Development.*: Ed. PACK Publishing.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. En R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver (Canada): Educational Research Institute of British Columbia.

Wingersky, M. S., Barton, M. A. y Lord, F. M. (1982). *LOGIST user's guide*. Princeton, New Jersey (USA): Educational Testing Service.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B. y Wesslén, A. (2000). *Experimentation in Software Engineering: An Introduction*. Massachusetts (EE.UU): Kluwer Academic Publishers.

Wolfe, J. H., McBride, J. R. y Sympson, J. B. (1997). Development of the experimental CAT-ASVAB system. En W. A. Sands, B. K. Waters y J. R. McBride (Eds.), *Computerized adaptive testing: from inquiry to operation* (pp. 97-101). Washington, DC (USA): American Psychological Association.

Wonnacott, R. J. y Wonnacott, T. H. (1991). *Estadística Básica Práctica. Su utilidad y múltiples aplicaciones*. México: LIMUSA.

Worthen, B. R., Sanders, J. R. y Fitzpatrick, J. L. (1997). *Program Evaluation. Alternative Approaches and Practical Guidelines,2nd ed*. New York (EE.UU.): Addison Wesley Longman.

Wright, B. D. y Bell, S. R. (1984). Items banks: what, why and how. *Journal of Educational Measurement, 21*(4), 331-346.

Wright, B. D. y Linacre, J. M. (1985). *MICROSCALE manual*. Westport, Conn (USA): Mediax Interactive Technologies, Inc.

Wright, B. D. y Linacre, J. M. (1992). *BIGSTEPS Rasch analysis*. Chicago, IL (USA): MESA Press.

Wright, B. D., Linacre, J. M. y Schulz, M. (1989). *BIGSCALE: Rasch analysis computer program*. Chicago, IL (USA): MESA Press.

Wright, B. D., Mead, R. J. y Bell, S. R. (1979). *BICAL: A Rasch program for the analysis of dichotomous data*. Chicago, IL (USA): MESA Press.

Wright, B. D. y Panchapakesan, N. A. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-37.

Zickar, M. J., Overton, R. C., Taylor, R. y Harms, H. J. (1999). The development of a computerized selection system for computarized programmers in a financial services company. En F. Drasgow y J. B. Olson-Buchanan (Eds.), *Innovations in computarized assessment* (pp. 7-34). Mahwah, New Jersey (USA): Lawrence Erlbaum Associates.

Zimowski, M. F., Muraki, E., Mislevy, R. J. y Bock, R. D. (1996). *BILOG-MG: multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL (USA): Scientific Software International.