Facultad de Filología
Departamento de Lengua Española

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

TESIS DOCTORAL

ESTUDIO DEL COMPORTAMIENTO DE LOS EXAMINADORES DE LA PRUEBA DE EXPRESIÓN ESCRITA MEDIANTE EL MODELO *MANY-FACET RASCH MEASUREMENT* (MFRM) EN EL CONTEXTO DE UN EXAMEN DE DOMINIO: EL DIPLOMA DE ESPAÑOL NIVEL A2

Juan Miguel Prieto Hernández
Directores: Juan Felipe García Santos y Gerardo Prieto Adánez

2016

VNiVERSiDAD D SALAMANCA

Facultad de Filología
Departamento de Lengua Española

ESTUDIO DEL COMPORTAMIENTO DE LOS EXAMINADORES DE LA PRUEBA DE EXPRESIÓN ESCRITA MEDIANTE EL MODELO *MANY-FACET RASCH MEASUREMENT* (MFRM) EN EL CONTEXTO DE UN EXAMEN DE DOMINIO: EL DIPLOMA DE ESPAÑOL NIVEL A2

Tesis para optar al grado de doctor
presentada por
Juan Miguel Prieto Hernández

V°B

Dr. Juan Felipe García Santos
Director

V°B°

Dr. Gerardo Prieto Adánez
Director

Salamanca, 2016

AGRADECIMIENTOS

Durante la realización de este trabajo he recibido el apoyo de numerosas personas. Vaya mi agradecimiento a todos ellos, y de manera muy especial a los que siguen.

Desde el comienzo de esta investigación he tenido la suerte de contar con dos espléndidos maestros que me han dedicado generosamente su tiempo y su saber: los catedráticos Juan Felipe García Santos y Gerardo Prieto Adánez. Sin ellos, sin sus ánimos y sus consejos, este trabajo no habría sido posible.

Me gustaría mostrar mi agradecimiento a todos los directores de Cursos Internacionales que han confiado en mí y me han encomendado tareas de responsabilidad en la evaluación de segundas lenguas: José Jesús Gómez Asencio, Juan Felipe García Santos, Jesús Fernández González, Noemí Domínguez García y José Miguel Sánchez Llorente. También debo agradecer a Mª Ángeles Pérez López el apoyo que me ofreció para que pudiera conseguir los datos necesarios para la redacción de esta tesis. Hago extensivo este agradecimiento a Francisco Moreno Fernández, Director académico del Instituto Cervantes en el momento de gestación de este trabajo, las facilidades que me brindó para que el proyecto pudiera echar a andar, y a mi compañero en Cursos Internacionales Luis Lucas Postigo, que tuvo que trabajar duro para facilitarme los datos que necesitaba.

No puedo olvidarme en esta relación de agradecimientos de mis actuales compañeros del Área de evaluación de Cursos Internacionales, a todos, a los de un lado y otro del Patio de Escuelas Menores, con los que he compartido los avances en mi trabajo. Debo mencionar de manera especial a Daniel Escandell Montiel, quien me ha ayudado con los aspectos más abstrusos de la bibliografía. Lógicamente, cualquier error en la misma ha sido problema de comprensión por mi parte; queda eximido de cualquier responsabilidad.

Debo hacer mención especialísima, cómo no, a mis compañeros y amigos del alma Marisol Martín Martín, Alberto Buitrago Jiménez y Elena Natal Prieto, sin cuyos ánimos, sin los cafés en los que hemos intercambiado conversaciones, quejas y proyectos y sin su apoyo absolutamente incondicional, mi trabajo no habría visto la luz.

Y también, a mis maestros de la Facultad de Filología, que me han apoyado y animado en todo momento. A los profesores Luis Santos Río y a Julio Borrego Nieto, por sus ánimos y por sus siempre sabios consejos; al profesor Jesús Fernández González y al profesor José Luis Herrero Ingelmo por su apoyo y por sus siempre acertadas recomendaciones en las charlas informales en los cafés del entorno de la Plaza de Anaya; y a todas las personas que, desde la distancia, siempre me han manifestado su apoyo.

Mi más profundo agradecimiento es para mi familia, sin la cual ni esto ni nada tendría sentido. A Charo, mi esposa, y a mis hijas, Laura y Clara, quienes siempre han confiado en mí. Gracias por su apoyo y constantes ánimos para que realizara este trabajo. Confío en estar a la altura de sus expectativas.

I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be.

William Thomson (Lord Kelvin), 1883.

ÍNDICE

I N T R O D U C C I Ó N

La acreditación del grado de competencia y dominio del idioma español comenzó a realizarse de manera regular y organizada considerablemente más tarde que en otras lenguas. Hasta finales de la década de los años 80 del siglo pasado no aparecieron las primeras propuestas de certificacion y, desde entonces, los esfuerzos por situar a la evaluación del español como segunda lengua en un nivel similar al de países de nuestro entorno han sido desarrollados, principalmente, por instituciones públicas. Es destacable la aportación que han hecho a este respecto el Instituto Cervantes y la Universidad de Salamanca.

La aparición en 2001 de las versiones en inglés y francés del *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación* (Consejo de Europa 2002b)[1] supuso una revolución, al adoptar tanto en el ámbito de la enseñanza como en el de la evaluación de lenguas un enfoque centrado en la acción, en la medida en que considera que los usuarios de la lengua y los alumnos que la aprenden actúan como miembros de una sociedad que tiene que llevar a cabo tareas, lingüísticas y no lingüísticas. La publicación del propio *Marco de referencia*, la abundante bibliografía que se editó durante su periodo de gestación, así como la que se generó tras su llegada, ha sido y es de inestimable ayuda para alumnos, profesores, formadores, creadores de métodos de enseñanza, evaluadores, investigadores… De hecho, la aparición del *Marco de referencia*, así como la de los diferentes niveles de referencia para las lenguas europeas que se han ido publicando siguiendo las recomendaciones que propuso el Consejo de Europa en el año 2001, ha provocado que la mayor parte de los sistemas de evaluación de Europa se replanteen tanto los procedimientos utilizados en sus exámenes para

[1] En adelante utilizaremos la denominación de *Marco de referencia*, tal y como el propio documento aconseja en las "Notas para el usuario del *Marco común europeo de referencia",* pp. XI-XIV.

determinar el grado de competencia y dominio, como la vinculación de los mismos con los niveles del *Marco de referencia*. En el caso del español, el Instituto Cervantes ha sido quien ha culminado el cumplimiento de la mencionada directriz al publicar el *Plan curricular del Instituto Cervantes. Niveles de referencia para el español* (en adelante, PCIC) (Instituto Cervantes 2007) —obra pionera en el ámbito europeo—, que presenta un amplio y detallado desarrollo de inventarios de especificaciones de la lengua española. Aunque en nuestro trabajo nos vamos a centrar en el análisis de la prueba de Expresión e interacción escritas (EIE), consideramos conveniente realizar una serie de consideraciones generales acerca de qué apartados de estos dos documentos es preciso tener en cuenta a la hora de diseñar y confeccionar las especificaciones de unos exámenes de dominio.

Los tests de desempeño o de ejecución (como es el caso de la prueba de EIE) son cada vez más habituales en la evaluación educativa. Sin embargo, su utilización no está exenta de polémica. Las calificaciones que se obtienen en este tipo de pruebas son objeto de crítica en muchas ocasiones, bien por no compartir la puntuación otorgada por el examinador, bien por no entender cómo se ha llegado hasta ella. Un ejemplo: es frecuente que entre los estudiantes que se presentan a las Pruebas de Acceso a las Enseñanzas Universitarias Oficiales de Grado se considere que la calificación obtenida en las pruebas de idiomas varía dependiendo del examinador que haya calificado dicha prueba: si es más severo, la puntuación será más baja, y si, por el contrario, es más benévolo, será más alta. Lógicamente, para que el proceso de evaluación se realice con todas las garantías, es necesario que los evaluadores dispongan de criterios y de escalas de calificación. Sin embargo, es habitual comprobar que, año tras año, a pesar existir unos criterios específicos de calificación, entre el amplio grupo de calificadores existen diferencias considerables a la hora de asignar las calificaciones. Los examinadores, por su parte, necesitarían saber si están siendo excesivamente rigurosos o indulgentes a la hora de interpretar los criterios y las bandas de

calificación. Sin información, el proceso puede no tener la suficiente fiabilidad.

El principal objetivo de esta investigación es analizar el comportamiento de los calificadores de la prueba de EIE de un examen de dominio: el Diploma de Español Nivel A2, del sistema de certificación de los Diplomas de Español como Lengua Extranjera (DELE). Al tratarse de títulos oficiales que otorga el Instituto Cervantes en nombre del Ministerio de Educación, Cultura y Deporte de España, las consecuencias de tener una nota u otra en el examen en general, o en una de las pruebas, por ejemplo, la de EIE, puede tener consecuencias relevantes: conseguir un puesto de trabajo, ascender profesionalmente, entrar o no en la universidad, conseguir una beca, etc.

Cualquier formador de formadores que imparta cursos dirigidos a examinadores de las pruebas de EIE es consciente de que parte de los asistentes a los cursos de formación aplica de manera muy diversa los criterios de calificación. Lo mismo sucede cuando un responsable de calificación de un sistema de certificación compara las calificaciones que los examinadores dan a los candidatos en una prueba de ejecución. Es habitual observar que hay examinadores que, por norma general, son menos exigentes, mientras que hay otros que son más severos. El sistema tradicional de calificación de la prueba de EIE en el ámbito de la certificación consistía en que cada candidato era calificado por dos examinadores y, posteriormente, se hallaba la media[2]. Si la diferencia era excesiva y para un examinador era apto en la prueba y para otro no, la prueba del candidato era calificada por un tercer examinador. Los calificadores estaban agrupados en parejas, y los dos miembros de cada pareja calificaban a los mismos candidatos. De este modo, cuando se comparaban las puntuaciones de los calificadores, era posible observar que había algunos que, aparentemente, eran, de manera sistemática, o más severos o más

[2] Así sucedía, por ejemplo, en los DELE hasta, precisamente, el comienzo de la redacción de esta tesis doctoral.

benévolos. Pero, severos o benévolos, ¿respecto a qué? El término de comparación siempre era su pareja calificadora. Pero ¿y si el que había sido severo o benévolo había sido el compañero?

Con el fin de que los datos obtenidos en los análisis que realizaremos en nuestro trabajo en relación con el comportamiento de los calificadores de la prueba de EIE permitan realizar comparaciones entre todo el equipo de calificadores, y no únicamente para cada pareja, proponemos establecer una red entre todos los calificadores, de manera que cada uno de ellos califique pruebas puntuadas por el resto. En el diseño de la red, seguiremos alguna de las opciones de distribución de las pruebas entre los examinadores que sugiere la bibliografía.

Para poder estudiar el comportamiento de los examinadores en el proceso de calificación de la prueba de EIE es preciso utilizar un modelo de análisis psicométrico que permita analizar de forma simultánea las diferentes variables que pueden tener un impacto relevante en los resultados de evaluación, y obtener medidas objetivas o invariantes. También es necesario que cada examinador puntúe pruebas de EIE calificadas por el resto de examinadores que participan en el proceso, de modo que sea posible comparar la severidad o benignidad de cada uno con todos los demás, y no solo con uno de ellos. Un modelo de análisis que permitirá realizar este estudio es *Many-Facet Rasch Measurement* (MFRM), descrito teóricamente por John Michael Linacre en 1989 y que dispone de un programa informático (FACETS) que facilita la realización de los cálculos que describe el modelo.

Por medio del modelo MFRM es posible analizar el modo o estilo de calificar de un grupo de examinadores y determinar si hay hábitos adquiridos o tendencias idiosincrásicas que afecten al proceso. FACETS permite visualizar el mapa de la variable en una única tabla en la que se presentan los elementos de las diferentes facetas analizadas calibrados en la misma escala de intervalos (*logit*). De este modo, es posible comparar e interpretar los resultados de la

competencia de los candidatos, la severidad de los calificadores, la dificultad de las tareas y de los atributos, así como determinar si las rúbricas o categorías de calificación se han utilizado de forma coherente y distinguible por el conjunto de los calificadores o por cada uno de ellos. En el primer caso, se emplea el Modelo de Escalas de Calificación (RSM); en el segundo, el Modelo de Crédito Parcial (PCM).

Asimismo, tenemos como objetivo profundizar en la explicación de la variabilidad que se produce en las actuaciones de los calificadores. Además de la severidad/benignidad, por medio del análisis que realizaremos podremos detectar otras disfunciones que pueden observarse en la actuación de los examinadores: tendencia central y efecto de halo. Por otro lado, además de a los calificadores, queremos analizar el resto de facetas que intervienen en el proceso de calificación: candidatos, atributos y tareas. Nuestra intención es demostrar que la aplicación de este sistema de medición en el área de conocimiento de la lingüística, especialmente en la aplicada, puede coadyuvar a que los procesos de calificación de las pruebas de desempeño o ejecución se realicen con mayores garantías de control y seguridad.

C A P Í T U L O 1

La evaluación de segundas lenguas

La evaluación está presente en múltiples ámbitos de la vida: en los estudios, en el deporte, en el trabajo… El hecho de examinar a alguien para saber cuánto sabe o lo que ha aprendido es algo que lleva haciéndose desde hace mucho tiempo, aunque bien es verdad que no siempre del mismo modo. En la antigüedad, por ejemplo, los maestros preguntaban oralmente a sus discípulos para saber si habían entendido bien sus explicaciones o no y en China, a partir del siglo VII, para acceder al funcionariado era preciso superar una serie de pruebas. Aquí, en Europa, tras la creación de las primeras universidades en los siglos XI y XII, los estudiantes, aunque no tenían que hacer exámenes en los cursos individuales, sí tenían que pasar exámenes orales ante tribunales académicos para obtener el grado.

Pero no fue hasta finales del siglo XVIII, en 1792, cuando un profesor de Química e Historia natural de la Universidad de Cambridge, Willian Farish, comenzó a calificar con notas los trabajos de sus estudiantes (Stray 2001). Algo más de medio siglo más tarde, en 1857, la Universidad de Oxford crea la University of Oxford Delegacy of Local Examinations (UODLE) con el propósito de realizar exámenes en centros próximos a los lugares donde vivían los candidatos. Las primeras pruebas se realizaron en julio del año siguiente. Ese mismo año, en 1858, la Universidad de Cambridge funda University of Cambridge Local Examinations Sindicate (UCLES). Los primeros exámenes se realizaron a finales de año, y 370 candidatos escolares fueron examinados en siete ciudades inglesas (Raban 2008, <http://www.cambridgeassessment.org.uk/about-us/who-we-are/our-heritage/>).

1.1. Los exámenes de dominio

En el ámbito de la evaluación de idiomas, entre las lenguas de nuestro entorno, el primer examen acreditativo del grado de competencia de una lengua europea para extranjeros del que se tiene noticia es el que realizó en 1913 la Universidad de Cambridge: el Certificate of Proficiency of English (CPE) como consecuencia del auge del estudio de las lenguas modernas que se vivió en Europa a comienzos del siglo XX. En la primera convocatoria únicamente se presentaron tres candidatos. Este examen, que en la actualidad equivale al nivel C2 del *Marco de referencia,* era en el momento de su creación un examen con el que se pretendía acreditar un excelente conocimiento de la lengua inglesa.

En 1934 el Reino Unido crea el British Committee for Relations with Other Countries, institución que dos años más tarde adoptó la denominación actual de British Council, con la misión de promover la educación, establecer relaciones culturales y crear vínculos internacionales (<http://www.britishcouncil.org/organisation>). Algo más de un cuarto de siglo más tarde del lanzamiento del primer examen para alumnos extranjeros, en 1939, la Universidad de Cambridge pone en circulación el segundo de los certificados con un nivel más bajo y dirigido a un público más amplio: el First Certificate in English (FCE), que en la actualidad equivale al nivel B2 del *Marco de referencia* (Raban 2008).

En la década de los años 40 las instituciones educativas de diversos países europeos comienzan a crear organismos con objetivos similares a los del British Council y, por lo general, su creación va acompañada del lanzamiento de exámenes certificativos de las lenguas que representan. El gobierno de Francia crea en 1945 el CIEP (la Alliance Française había sido creada en 1883 en torno a personalidades célebres como Julio Verne, Luis Pasteur, Ferdinand de Lesseps, Armand Colin o Jules Renan) y en 1985 el Ministerio de Educación francés publica un decreto ministerial por el que se crean los

exámenes del DELF y DALF (<http://www.ciep.fr/es>). Por su parte, el gobierno alemán crea en 1951 el Instituto Goethe con la misión de capacitar a profesores de alemán como lengua extranjera en Alemania (<https://www.goethe.de/en/uun/org/ges.html)>. En 1998, la Società Dante Alighieri, que había sido creada en 1889 por un grupo de intelectuales italianos con el propósito de proteger y difundir la lengua italiana por el mundo, implementó el certificado internacional del idioma italiano PLIDA (<http://www.danterosario.com.ar/AsocHistoria.es.php>). Finalmente, el gobierno portugués convirtió el *Instituto da Cultura e Língua Portuguesa* en el Instituto Camões en 1992 quien, junto con el Centro de Avaliação do Português Língua Estrangeira (CAPLE), creó el Sistema de Evaluación y Certificación del Portugués Lengua Extranjera (SACLEP) (<http://www.ulisboa.pt/wp-content/uploads/Despacho-n.%C2%BA-3305-2015.pdf>).

1.1.1. Diplomas de Español Como Lengua Extranjera (DELE)

En el caso del español fue la Universidad de Salamanca la que a través de su Servicio de Cursos Internacionales inició en 1987 el proceso de creación del Diploma de Español de la Universidad de Salamanca (DEUS), destinado a acreditar al estudiante extranjero en el nivel superior de dominio del español. Los exámenes para la obtención de este título se comenzaron a celebrar en 1988. El Boletín Oficial del Estado del 29 de julio de ese año publicó el Real Decreto 826/1988, mediante el que se creaban los diplomas acreditativos del conocimiento del español como lengua extranjera (niveles básico y superior), en el que se indicaba que el Ministerio de Educación y Ciencia elaboraría las pruebas de examen para la obtención de los Diplomas de Español como Lengua Extranjera (en adelante DELE) (<http://www.boe.es/diario_boe/txt.php?id=BOE-A-1988-18767>). La

primera convocatoria se realizó en noviembre de 1989 y fue coordinada por la Subsecretaría de Cooperación Internacional.
En 1990 la Universidad de Salamanca realiza la primera convocatoria del Certificado de Español de la Universidad de Salamanca (CEUS), que acreditaba al estudiante extranjero un nivel intermedio de dominio del español.

A finales de 1989, la Universidad de Cambridge y la Universidad de Salamanca acordaron crear una asociación de entidades examinadoras de lenguas europeas, la Association of Language Testers in Europe (ALTE), que en la actualidad cuenta con 34 miembros de todos los países de la Unión Europea, que representan 26 lenguas[3]. Desde sus inicios, uno de los principales objetivos de ALTE ha sido el de establecer escalas de niveles comunes de competencia con el fin de facilitar el reconocimiento transnacional de las diversas certificaciones otorgadas por las diferentes instituciones certificadoras europeas.

En febrero de 1991 el Ministerio de Educación y Ciencia y la Universidad de Salamanca firman un convenio en virtud del cual los exámenes son elaborados por la Universidad de Salamanca y esta asume la responsabilidad científica de crear y evaluar los exámenes para la obtención de ambos Diplomas. En el mes de junio se realiza la primera convocatoria de exámenes conjunta entre el Ministerio de Educación y la Universidad de Salamanca.

El Instituto Cervantes se crea en ese mismo año (BOE del viernes 22 de marzo de 1991, ley 7/1991) con el mandato expreso de "organizar las pruebas de verificación del conocimiento del Español, para la obtención de los diplomas oficiales expedidos por el MEC, en los términos que éste regule".

[3] Datos consultados en www.alte.org el 30 de noviembre de 2015.

A partir de esta fecha la Universidad de Salamanca viene colaborando con el Instituto Cervantes en la elaboración y calificación de las pruebas de los diferentes niveles de los DELE. Desde la creación del Instituto, dos fechas clave en su trayectoria en relación con la certificación son: (i) la publicación el 8 de noviembre de 2002 en el BOE del Real Decreto 1137/2002, de 31 de octubre, en virtud del cual el Ministerio de Educación transfiere los DELE al Instituto Cervantes, y (ii) el Real Decreto 264/2008, de 22 de febrero, y publicado el 12 de marzo por el que se cambia la denominación de los diplomas y estos pasan a vincularse con la escala de niveles del *Marco de referencia* (pueden consultarse ambos Reales Decretos en <http://diplomas.cervantes.es/informacion/descripcion_dele.html>).

1.1.2. Linguaskill

En 1994 la multinacional de servicios de empleo Manpower se dirigió a la Universidad de Cambridge para solicitar un examen adaptativo por ordenador para poder conocer de manera rápida (como máximo en una hora de duración) el nivel de dominio lingüístico de usuarios en inglés, francés, alemán, español, italiano y holandés (Jones 2014: 63-65).

En una primera fase el examen se diseñó en inglés y posteriormente se desarrolló en alemán, español y francés con la colaboración de los siguientes miembros de ALTE: Instituto Goethe, Universidad de Salamanca y Alianza Francesa. Con posterioridad se incorporó el holandés con la colaboración de la Universidad Católica de Lovaina, desarrolló el proyecto Linguaskill, un examen adaptativo por ordenador (Computer Adaptive testing –CAT, en sus siglas en inglés–)

El modelo elegido fue un examen adaptativo computerizado (CAT) denominado Linguaskill que evalúa las cuatro habilidades lingüísticas:

comprensión auditiva y de lectura, expresión oral y escrita que se adapta de forma automática al nivel del usuario. Al finalizar la prueba se informa al candidato de la puntuación que ha alcanzado según la escala de niveles de ALTE. Los ítems que configuran los diversos exámenes se seleccionan automáticamente para cada usuario dependiendo de su nivel y están vinculados con la escalas de niveles del *Marco de referencia* mediante el modelo de Rasch (Geranpayeh, Ardeshir 2001b, 2001c).

En el certificado que se expide al finalizar el examen se informa al usuario del nivel alcanzado en una escala de 0 a 100, en la que el 0 se situaría en el nivel más bajo de la escala y 100 en el más alto. Este valor es una modificación o aclaración de los valores escalares de los candidatos y de los ítems, que habitualmente se expresan en la escala *logit*. La escala, denominada Level Description, se subdivide a su vez en 6 bloques que se corresponden con los niveles de la escala de ALTE y del *Marco de referencia* (cf. Tabla 4). También se informa al usuario de lo que puede ser capaz de hacer lingüísticamente, según el nivel alcanzado.

La primera versión de Linguaskill se desarrolló para el sistema operativo DOS que fue utilizada por Manpower durante un año y en 1996 fue sustituida por una versión para Windows.

1.1.3. CommuniCAT

Tras el éxito de Liguaskill se desarrollaron diversos proyectos de exámenes adaptativos en colaboración con miembros de ALTE. Seis miembros de ALTE: Certificaat Nederlands als Vreemde Tall (holandés), UCLES (inglés), Alianza francesa (francés), Guethe-Institut (alemán), Università per Stranieri di Perugia (italiano) y Universidad de Salamanca (español) constituyeron el 17 de noviembre de 1998 el grupo KoBalt (Komputer Based Language Testing)

con el objetivo de establecer una alianza estratégica dentro de Europa para desarrollar y potenciar la innovación en el ámbito de las pruebas de idiomas. Consecuencia de esta colaboración son otros exámenes adaptativos por ordenador, como ComuniCAT, destinado al ámbito de los negocios (Jones 2014: 65).

1.1.4. BULATS

BULATS es uno de los productos de la familia CommuniCAT. Se trata de un sistema de exámenes centrado en el ámbito laboral que permite a empresas y trabajadores tomar decisiones relacionadas con la formación, contratación, promoción o movilidad de los trabajadores. BULATS se ha desarrollado en cuatro idiomas (inglés, francés, alemán y español) y permite determinar de forma rápida y fiable (Geranpayeh 2001a) el nivel de dominio lingüístico de cualquier usuario en relación con la escala de seis niveles del *Marco de referencia* (tabla 1). En el examen BULATS el sistema determina el lugar que el usuario ocupa en dicha escala.

Una de las versiones de BULATS, el examen por ordenador, está disponible on line y en CD-ROM y se evalúan las destrezas de comprensión oral y escrita, el conocimiento de la lengua y el vocabulario. Esta versión de BULATS, al igual que Linguaskill, es un examen adaptativo que se ajusta al nivel del candidato, de forma que dependiendo de la respuesta que el candidato dé a cada pregunta, el sistema seleccionará preguntas más fáciles o difíciles con el fin de adaptarse a su nivel y no ofrecerle preguntas excesivamente sencillas o complicadas.

BULATS también se comercializa en otras versiones:

- Examen estándar: se realiza en papel y se evalúan las destrezas de comprensión oral y escrita, el conocimiento de la lengua y el vocabulario.
- Examen de expresión escrita: se realiza en papel y se evalúa la destreza de expresión escrita.
- Examen on line de expresión escrita: está disponible on line y se evalúa la destreza de expresión escrita. Sólo está disponible en inglés.
- Examen de expresión oral: se realiza en papel y se evalúa la destreza de expresión oral.
- Examen on line de expresión oral: está disponible on line y se evalúa la destreza de expresión oral; las respuestas se grabarán mediante un micrófono. Sólo está disponible en inglés.

El 26 de abril de 2012 se presentó CertiUni, la plataforma de certificación universitaria promovida por la Conferencia de Rectores de las universidades españolas (CRUE), con el objetivo de certificar a alumnos y profesores de la universidad española sus competencias en las áreas de informática, idiomas y competencias personales que especifica el Espacio Europeo de Educación Superior (EEES) (<http://www.eees.es/es/eees>) y que recoge la Declaración de Bolonia del 19 de junio de 1999 (<http://www.certiuni-crue.org/>).

En el ámbito de la certificación de lenguas, la Plataforma de Certificación Universitaria ha seleccionado BULATS para que los universitarios españoles puedan acreditar su nivel de conocimientos lingüísticos. En virtud de un acuerdo entre la CRUE y la Universidad de Salamanca en 2010 (Memoria de Cursos Internacionales de la Universidad de Salamanca, S.A. 2010-2011, <http://cursosinternacionales.usal.es/index.php/es/memorias-de-actividades/item/download/8>), Cursos Internacionales de la Universidad de Salamanca se ha convertido "en el único suministrador para los exámenes que a través de este medio se desarrollen en las Universidades españolas" (Informe al Claustro del Rector de la Universidad de Salamanca, sesión del 15 de mayo de

2012, en <http://saladeprensa.usal.es/webusal/files/Informe_%20 Rector_ Claustro.pdf>).

El futuro del proyecto es incierto. A partir de finales de 2016 se reconsiderará la continuidad del consorcio Kobalt y cada una de las instituciones se replanteará la gestión del examen BULATS correspondiente a su lengua.

1.1.5. SurveyLang

En la reunión del Consejo Europeo celebrado en Barcelona en 2002, los Jefes de Estado y de Gobierno solicitaron el establecimiento de un *Indicador europeo de competencia lingüística* (en adelante *Indicador*) con el fin de proporcionar información a los Estados miembros de la situación de la enseñanza-aprendizaje de idiomas en sus respectivos países. El 1 de agosto de 2005 la Comisión Europea presentó una comunicación en la que se describe el *Indicador* y se detallan sus parámetros generales, el planteamiento de recogida de datos y aspectos relacionados con la gestión. (<http://eurlex.europa.eu/smartapi/cgi/sga_doc?smartapi!celexplus!prod! DocNumber&lg=es&type_doc=COMfinal&an_doc=2005&nu_doc=356>) La comunicación destaca la "necesidad de contar con una herramienta que permitiera a los gobiernos comprender los niveles existentes de dominio de segundas lenguas y hacer comparaciones significativas con otros países. Además, proporcionaría información sobre la influencia de las variables demográficas, sociales, económicas y educativas en el dominio lingüístico tanto dentro de cada país como entre los estados miembros" (<http://www.surveylang.org/es/About-SurveyLang/A-brief-history.html>).

Dos años más tarde, el 13 de abril de 2007, en una comunicación de la Comisión al Consejo titulada *Marco para la encuesta europea sobre los conocimientos lingüísticos* (<http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:

2007:0184:FIN:ES:PDF>), se proporcionaban más detalles para la realización de la encuesta. Ese mismo año, tras la aceptación de este documento, la Comisión Europea abrió un concurso público para realizar la Encuesta. Finalmente, en febrero de 2008 se adjudicó el contrato al consorcio internacional SurveyLang, formado por ocho organizaciones: el Centro internacional de estudios pedagógicos (CIEP), Gallup Europa, el Instituto Goethe, el Instituto Cervantes, el Instituto Nacional para la medición en educación (CITO), la Universidad de Cambridge, la Universidad de Salamanca y la Universidad para Extranjeros de Peruggia.

La Encuesta Europea de Competencia Lingüística (ESCL) puede considerarse como un proyecto pionero, ya que es la primera vez que en Europa un consorcio de las instituciones más relevantes en la elaboración de pruebas de dominio de las lenguas más estudiadas en Europa, desarrolla unos tests lingüísticos con estándares comunes que permiten medir el dominio lingüístico de forma coherente y comparable entre las cinco lenguas más estudiadas en Europa: el inglés, el alemán, el francés, el español y el italiano. Los países que participan en la encuesta y las lenguas analizadas son: Bélgica (Comunidad Flamenca): francés, inglés; Bélgica (Comunidad Francesa): inglés, alemán; Bélgica (Comunidad Alemana): francés, inglés; Bulgaria: inglés, alemán; Croacia: inglés, alemán; Inglaterra: francés, alemán; Estonia: inglés, alemán; Francia: inglés, español; Grecia: inglés, francés; Malta: inglés, italiano; Holanda: inglés, alemán; Polonia: inglés, alemán; Portugal: inglés, francés; Eslovenia: inglés, alemán; España: inglés, francés y Suecia: inglés, español. Para alinear, por ejemplo, la expresión escrita en los cinco idiomas, los calificadores tuvieron que calificar las muestras de actuación en dos lenguas en las que eran competentes. Para el análisis de los datos se utilizó un procedimiento descrito en 2008 por Gilles Breton, Sylvie Lepage y Brian North y en 2009 por Neil Jones.

Los resultados de las pruebas de la Encuesta están vinculados con el *Marco de referencia*. Sin embargo, aunque el Consejo de Europa (2009a) ha publicado el *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR)* (en adelante *Manual*) no resulta especialmente útil para la Encuesta ya que en el *Manual* no se trata de forma explícita la dimensión multilingüe, que es fundamental para el proyecto del consorcio europeo. Para confirmar la vinculación de la Encuesta con el *Marco de referencia* se han utilizado las descripciones statements» («Puede hacer») de ALTE (vid. Abajo) en el cuestionario de los alumnos con el fin de obtener una autoevaluación de la competencia en las cuatro habilidades. Recordemos que Neil Jones había demostrado en 2002 la utilidad de las vinculaciones de los Can-Do con el *Marco de referencia*. Para el análisis de las pruebas se utilizó el Modelo Logístico de un parámetro (modelo de Rasch) descrito en 1995 por Verhelst y Glas y por Verhelst, Glas y Verstralen.

La encuesta evalúa el nivel de dominio de los alumnos en comprensión de lectura, comprensión auditiva y expresión escrita de las dos lenguas europeas (entre inglés, francés, alemán, italiano o español) más enseñadas en cada país en una muestra representativa de estudiantes de Enseñanza Secundaria y del primer curso de Bachillerato de más de 1.500 estudiantes por lengua evaluada y país. Para desarrollar el test el consorcio realizó los siguientes trabajos.

- Definición del constructo del test.
- Redacción de los ítems del examen.
- Experimentación.
- Determinación de las tareas que se usarán en el ensayo de campo y en la prueba principal.

Y para diseñar la encuesta se ha elaborado material específico:

- Cuestionarios para estudiantes, profesores, directores de escuelas y los coordinadores nacionales con el fin de obtener información contextual adicional para complementar los análisis que se realicen.

- Sistemas de herramientas informáticas interconectadas para las diferentes fases del proceso:

    - Herramienta de autoría/creación de ítems (Item Authoring Tool): permite a los redactores de ítems trabajar de forma independiente por medio de plantillas para cada uno de los tipos de tareas.
    - Banco de ítems de examen (Test Item Databank): se trata de un Banco de Items de Examen para almacenar los ítems de examen.
    - Herramienta de construcción de examen (Test Assembly Tool); mediante información de cada estudiante crea exámenes personalizados.
    - Herramienta de control de examen (Test Rendering Tool), permite realizar el examen en un ordenador por medio de una memoria USB o un CD de manera que SurveyLang controla el escritorio y bloquea ciertos recursos del ordenador, como el acceso a internet.

Como continuación del proyecto SurveyLang, en 2015 la Comisión europea sacó a licitación un proyecto denominado "Study on comparability of language testing in Europe" (<http://ec.europa.eu/education/calls/tender-45-2014_en.htm>) que consiste en realizar un meta-análisis de los exámenes nacionales de idiomas que existen en los diferentes Estados miembros (como la PAU en España, el baccalaureat en Francia, etc.) para determinar hasta qué punto son comparables los resultados obtenidos. La licitación fue ganada por ALTE. Para la realización del estudio se seguirán las siguientes fases:

- Recogida de datos y ejemplos de los diferentes exámenes de los 28 Estados miembros.

- Analizar los exámenes de los diferentes idiomas con un instrumento de medida standard (una lista con preguntas, similar a las checklists de ALTE pero mas breve), de forma que sea posible sistematizar la información de todos los exámenes y que luego se puedan hacer comparaciones.
- Redactar los informes en los que se resuman los resultados del análisis, con recomendaciones para los distintos países sobre cómo mejorar sus exámenes nacionales de idiomas de forma que los datos obtenidos sean mas comparables entre países.

1.1.6. SIELE

En el III Congreso Internacional de la Lengua Española celebrado en Rosario (Argentina) en noviembre de 2004, se emplazó a las instituciones hispanohablantes a la creación de un sistema de certificación destinado a los estudiantes de español. Siete meses más tarde, en junio de 2005, se celebraron en Rosario, las Jornadas para la Certificación Unitaria del Español como Lengua Extranjera, en las se continuó debatiendo el proyecto. En octubre de 2005, en la Reunión de Rectores que tuvo lugar en Salamanca (España) con motivo de la Cumbre Iberoamericana de Jefes de Estado y de Gobierno, se acordó la creación de un sistema de certificación internacional del español como lengua extranjera. En las jornadas previas al IV Congreso Internacional de la Lengua Española, celebrado en Cartagena de Indias (Colombia) en marzo de 2007, los rectores y representantes de más de un centenar de universidades del mundo hispanohablante ratificaron la constitución del Sistema Internacional de Certificación del Español como Lengua Extranjera (SICELE). El 2 de junio de 2010 se suscribió en Guadalajara (México) el Convenio Marco Multilateral SICELE, lo que supuso la consolidación del SICELE como asociación internacional. Se adhirieron al convenio más de un centenar de instituciones universitarias (<http://www.sicele.org/>).

Desde aquella fecha se han celebrado dos congresos internacionales SICELE. El primero de ellos fue el Congreso Internacional Investigación e Innovación en ELE: construyendo el SICELE. "La investigación e innovación en la evaluación y divulgación del español como lengua extranjera en el siglo XXI", celebrado en Puerto Rico los días 9 al 11 de mayo de 2012 y organizado por la Universidad de Puerto Rico, la Universidad Interamericana y la Universidad del Sagrado Corazón. La información sobre este Congreso se encuentra disponible en la página electrónica <http://www.cisicelepr2012.org>

El segundo de los congresos, organizado por la Universidad Nacional Autónoma de México en colaboración con otras entidades académicas de Hispanoamérica, llevaba por título: II Congreso Internacional del SICELE "Unidad y diversidad del Español: Avances en la Certificación de E/LE", celebrado en Ciudad de México del 24 al 26 de noviembre de 2014. En este segundo congreso se dieron a conocer los avances del SICELE en relación con la definición de estándares de calidad y la promoción para la acreditación de diversos instrumentos de evaluación (<http://www.sicele.cepe.unam.mx/>)

El 2 de julio de 2015 los Reyes de España presidieron la presentación, también en Ciudad de México, del Servicio Internacional de Evaluación de la Lengua Española (SIELE), un examen que, aunque en una primera fase será progresivo, tiene como objetivo llegar a ser adaptativo y que permitirá determinar en muy poco tiempo el grado de conocimiento del español de cualquier persona, desde el nivel A1 hasta el C1 (cf. el discurso de Felipe VI en acto de presentación en <http://www.casareal.es/ES/Actividades/Paginas/actividades_discursos_detalle.aspx?data=5496>).

La iniciativa ha sido desarrollada conjuntamente por el Instituto Cervantes, la Universidad Nacional Autónoma de México y la Universidad de Salamanca. El examen, que integrará variantes de todas las zonas lingüísticas

de habla hispana, constará de cuatro pruebas: Comprensión lectora y Comprensión auditiva, cuya calificación será inmediata, y Expresión e Interacción oral y escrita, que serán evaluadas mediante escalas de calificación por examinadores acreditados. El certificado que se obtenga tendrá una validez de dos años (<www.siele.org>).

El 3 de septiembre de 2015 el director del Instituto Cervantes, Víctor García de la Concha; los rectores de la Universidad Nacional Autónoma de México (UNAM), José Narro Robles, y de la Universidad de Salamanca, Daniel Hernández Ruipérez, y el presidente de Telefónica, César Alierta Izuel, presentaron en la sede central del Instituto Cervantes el SIELE. En el acto también intervinieron los ministros de Asuntos Exteriores, José Manuel García-Margallo, y de Educación, Cultura y Deporte, Íñigo Méndez de Vigo. En el mismo acto se procedió además a la firma del protocolo para el comienzo de la colaboración con Telefónica Educación Digital, la empresa encargada del desarrollo tecnológico y la comercialización del Servicio (cf. <http://www.cervantes.es/sobre_instituto_cervantes/prensa/2015/noticias/siele-presentacion-madrid.htm>).

El SIELE comenzó a funcionar a comienzos de 2016 y se prevé que alcance 300.000 candidatos en una primera fase, que se elevarán a 750.000 en el plazo de cinco años. La previsión es que haya centros de examen en los cinco continentes, pero durante los tres primeros años se implantará especialmente en tres países: Brasil, donde está previsto que en ese periodo de tiempo haya 120 centros de examen que cubran el 81% del territorio; Estados Unidos, donde existirán 100 centros de examen con una cobertura del 70%; y China, con 60 centros de examen y cobertura del 61%.

1.2. Terminología

La terminología utilizada en español para denominar la evaluaciones que se realizan en relación con el aprendizaje de segundas lenguas es diversa y como

la competencia lingüística de un candidato puede medirse desde diferentes enfoques: desde el punto de vista del dominio lingüístico, de la aptitud, del nivel…, en cada una de estas denominaciones subyace una determinada intención o finalidad que conviene tener presente.

Una *evaluación de dominio*, tal y como describe el *Marco de referencia*, consiste en determinar lo que un candidato "sabe o es capaz de hacer en cuanto a la aplicación en el mundo real de lo que ha aprendido; representa, por tanto, una perspectiva externa" (183). Frente a este tipo de evaluación, el *Marco de referencia* sitúa la *evaluación del aprovechamiento*, que es la que se realiza en el contexto docente, en una de las fases del proceso de enseñanza aprendizaje y, por ese motivo, representa una perspectiva interna. Es decir, mientras que la *evaluación de dominio* trata de evaluar el dominio lingüístico de manera externa e independientemente de cualquier proceso de enseñanza aprendizaje, lo que persigue una *evaluación del aprovechamiento* es determinar el grado de cumplimiento de los objetivos de un curso.

Esta misma interpretación figura en el *Multilingual Glossary of Language Testing Terms* (ALTE 1999) (en adelante, *Glosario multilingüe*) en el que se define la *evaluación de dominio* como "prueba que mide la capacidad o destreza general, sin referirse a ningún curso o conjunto de materias en particular" así como en la traducción que hace Neus Figueras del libro de Alderson, Clapham y Wall, *Exámenes de idiomas. Elaboración y evaluación* (1998: 16 y 26).

También en la definición de las destrezas que evalúa la Encuesta Europea de Competencias Lingüísticas que realizo el consorcio SurveyLang (<http://www.surveylang.org/es/>), se afirma que "la encuesta evaluará el *nivel de dominio* [el subrayado es nuestro] de los estudiantes en comprensión auditiva, comprensión de lectura y expresión escrita."

Sin embargo, parece que no hay unanimidad en la utilización de la expresión *evaluación del dominio*. En la misma traducción de *Exámenes de idiomas*, al

comienzo del libro, en la Génesis y prólogo, al referirse a una invitación de uno de los autores del libro por Karl Kranhnke y Charles Stansfield para colaborar como redactor en la publicación por parte de TESOL de los *Reviews of English Language Proficiency Tests* (Alderson, Krahnke y Stansfield 1987), la traductora del libro traduce el trabajo como "Informes de los exámenes de *aptitud* [el subrayado es nuestro] en lengua inglesa", mientras que unas páginas más adelante, tal y como indicábamos antes, traduce el término *proficiency* como 'dominio'. También encontramos confusión entre ambos términos en el estudio de José Luis Rodríguez y Francisco Javier Tejedor titulado *Evaluación educativa: 1. Evaluación de los aprendizajes de los alumnos* (1996) en el que declaran que "las pruebas de dominio son aquellas que se usan en intervalos frecuentes en la enseñanza para evaluar y guiar el progreso del estudiante", interpretación que coincide, en lo fundamental, con la que hace el *Marco de referencia* para la *prueba de aprovechamiento*.

Otro de los términos usados para hacer referencia a un tipo de evaluación es el de evaluación de la proficiencia. Este anglicismo, que figura en el *Diccionario de términos claves de ELE* del Centro Virtual Cervantes (http://cvc.cervantes.es/Ensenanza/biblioteca_ele/diccio_ele/default.htm) (Instituto Cervantes, s.f.), se equipara en esta publicación electrónica con la de dominio y se define del siguiente modo:

> Se entiende por prueba de proficiencia […] la que tiene como finalidad determinar los conocimientos lingüísticos de un candidato y las posibilidades que tiene de desenvolverse en el mundo real, aplicando lo que sabe. (*Diccionario de términos claves de ELE*, bajo la entrada *proficiencia*)

Es decir, que la definición del vocablo coincide, en lo fundamental, con la que ofrece el *Marco de referencia* para la *evaluación del dominio*. Pero además, la contraposición que encontrábamos en el *Marco de referencia* entre *evaluación del dominio* y *evaluación del aprovechamiento*, la hallamos en el *Diccionario de términos clave de ELE* entre *evaluación de la proficiencia* y *evaluación del rendimiento*. Para los

autores del diccionario, en consecuencia, la evaluación de la proficiencia es equivalente a la de dominio y la del rendimiento a la del aprovechamiento.

No obstante, parece poco recomendable la utilización del vocablo 'proficiencia'. En primer lugar, el término no aparece en la última edición del *Diccionario de la Real Academia* (2014). Además, la equivalencia en español que ofrece el *Glosario multilingüe* para el término inglés 'proficiency' es 'dominio'. Y, finalmente, el propio *Diccionario de términos clave de ELE*, en la definición de *prueba de aptitud* (que se estudiará a continuación), señala que la expresión *prueba de proficiencia* es una "traducción al castellano poco afortunada del término inglés *proficiency*".

En el *Glosario multilingüe* encontramos la siguiente definición para *prueba de aptitud*: "prueba diseñada para la predicción o medición del potencial de éxito del candidato en un área específica del aprendizaje, por ejemplo, el aprendizaje de una lengua extranjera". Es decir, que según el glosario, el objetivo primordial de una *prueba de aptitud* es predecir o determinar el éxito potencial de un candidato. En esta línea, Alderson, Clapham y Wall (1998: 26) afirman que una *prueba de aptitud* (*aptitude test* en inglés) consiste en "hacer primero un análisis de las situaciones y del uso de la lengua meta, así como de las actuaciones lingüísticas que el propio examen prevé". La *aptitud*, en consecuencia, y tal y como la define el DRAE, es la "capacidad para operar competentemente en una determinada actividad". Una prueba de aptitud tiene una finalidad predictiva, de estimación de las posibilidades de éxito que tiene un candidato que supera la prueba.

El último tipo de prueba al que nos vamos a referir es el de *prueba de nivel*. Aunque en el *Glosario multilingüe* no figura esta expresión, sí se define el término *nivel*:

> Grado de dominio exigido para que un alumno/a acceda a una clase determinada o el grado que representa un examen determinado a menudo se expresa mediante una serie de niveles. Por lo general se les designa con

> nombres como 'elemental', 'intermedio', 'avanzado', etc. (*Glosario multilingüe,* bajo la entrada *nivel*)

Esta definición es similar a la que el propio *Glosario multilingüe* hace de *prueba de clasificación*: "prueba administrada para colocar a los alumnos en un grupo o clase que tenga el nivel apropiado según sus conocimientos y capacidades". Y también está en la misma línea de la definición que ofrece el *Diccionario de términos claves de ELE* para *prueba de nivel*: "[prueba] que tiene como finalidad establecer el nivel de lengua de los estudiantes que la realizan, para poder clasificarlos en grupos, lo más homogéneos posible, con el fin de que cada uno reciba la instrucción adecuada al nivel demostrado".

En consecuencia, el término que mejor se ajusta al tipo de evaluación que realizan los exámenes de los DELE es el de *evaluación de dominio*.

En nuestro trabajo utilizaremos el término *prueba* para referirnos a cada uno de los componentes principales de los que consta un examen. Así, un *examen* estará formado por diversas pruebas que evalúan habilidades o destrezas específicas.

C A P Í T U L O 2

La determinación del nivel lingüístico a partir de 1950

2.1. La certificación de los niveles lingüísticos en Estados Unidos a partir de 1950

A mediados del pasado siglo no existía ni en Estados Unidos ni en Europa tradición en la definición de niveles de competencia lingüística. En Estados Unidos, el conflicto bélico contra Japón que había comenzado en 1941 con el ataque de la aviación japonesa a la base naval de Pearl Harbor en Hawái y que finalizó con la rendición de Japón el 14 de agosto de 1945, así como la posterior guerra de Estados Unidos contra Corea del Norte —que contaba con el apoyo del ejército chino— entre 1950-1953 tras la división que se realizó de la antigua Corea al finalizar la Segunda Guerra Mundial, evidenció la falta de preparación en idiomas extranjeros de la población estadounidense. En 1952 la Comisión de Administraciones Públicas (Civil Service Commission) de Estados Unidos intentó elaborar un inventario con el fin de registrar la capacidad lingüística de los empleados del gobierno estadounidense. Los problemas con los que se encontró la comisión fueron, por un lado, el de la carencia de pruebas de certificación y, por otro, el de la inexistencia de una definición precisa de niveles lingüísticos. No era fácil, en consecuencia, ni para las autoridades norteamericanas ni para los propios trabajadores del Servicio Exterior estadounidense indicar lo *mal*, *regular*, *bien* o *muy bien* que se utilizaban las lenguas extranjeras (Herzog 2011). Este problema también se percibía desde el ámbito docente: en 1954, Willian Parker, presidente aquel año de la Asociación Americana de Lenguas Modernas (MLA, en sus siglas en inglés), en la primera edición su obra *The*

National Interest and Foreingn Languages (Parker 1957), "subrayaba la necesidad de estudiar lenguas por razones de defensa nacional" (Paz 1988: 27).

El Gobierno de los Estados Unidos se puso manos a la obra y, a través del Instituto de Servicios Exteriores Norteamericano (<http://www.state.gov/m/fsi/>, FSI, en sus siglas en inglés) comenzó a diseñar un sistema objetivo de definición de niveles que fuera aplicable a todos los idiomas e ideó una escala de seis niveles que iba de 1 a 6 y que no diferenciaba entre las cuatro destrezas. La utilización de esta escala entre los funcionarios del servicio exterior de Estados Unidos puso en evidencia las enormes carencias que había en relación con el aprendizaje de lenguas. Una de las consecuencias de este trabajo fue que en 1958, a los pocos años de la creación de dicha escala, las pruebas de competencia lingüística se convirtieron en obligatorias para todos los funcionarios del servicio exterior aunque bien es verdad que al no estar descritos los niveles de manera rigurosa y precisa —analizados desde una perspectiva actual—, las pruebas, que estaban basadas en estas escalas, adolecían de una alta subjetividad. Con posterioridad se revisó la escala en diversas ocasiones y finalmente quedó estandarizada en seis niveles, aunque se modificó la numeración y se denominó 0 al más bajo (sin capacidad funcional) y 5 al más alto (nivel equivalente al de un hablante nativo).

Vinculada a esta nueva escala, la FSI diseño una prueba de expresión oral, con el formato de una entrevista que fue adoptada por otras muchas agencias del gobierno. En 1968 diversas agencias de cooperación estadounidenses redactaron descripciones para cada uno de los niveles y las cuatro destrezas. La consecuencia fue que la escala resultante comenzó a ser utilizada por el gobierno de los Estados Unidos para medir el nivel de competencia lingüística de todo su personal. La Organización del Tratado del Atlántico Norte (OTAN) adoptó en 1976 como escala de competencia lingüística una versión relacionada con esta prueba de Expresión oral. En la década de los 80, el Consejo Americano para la Enseñanza de Lenguas Extranjeras

(ACTFL, en sus siglas en inglés) desarrolló y publicó unas nuevas escalas de nivel que, aunque se basaban en las de la FSI, reducían el número de niveles a cuatro: superior, avanzado, intermedio y principiante.

Tabla 1
Equivalencias entre la escala FSI y la de ACTFL

Escala del Gobierno estadounidense	Escala de ACTFL
5	Superior
4	—
3	—
2	Avanzado
1	Intermedio
0	Principiante

Basada en la tabla 4 de Swender 2001.

2.2. El *Marco de referencia*

El Consejo de Europa comenzó a gestarse tras la finalización de la Segunda Guerra Mundial. Algo más de un año después de que finalizara el conflicto bélico, Winston Churchil, en un discurso pronunciado en la Universidad de Zurich, realizó un llamamiento a favor de la creación de unos Estados Unidos de Europa. Tras una serie de hitos en el proceso de creación del mencionado organismo —como el congreso celebrado en La Haya en mayo de 1948— el 5 de mayo de 1949, diez países (Reino de Bélgica, Reino de Dinamarca, República Francesa, República Irlandesa, República Italiana, Gran Ducado de Luxemburgo, Reino de los Países Bajos, Reino de Noruega, Reino de Suecia y Reino de Gran Bretaña y de Irlanda del Norte) firmaron en Londres el tratado por el que se fundaba el Consejo de Europa con la finalidad de "realizar una unión más estrecha entre sus miembros para salvaguardar y promover los ideales y los principios que constituyen su patrimonio común y favorecer su progreso económico y social" (Instrumento de Adhesión de España… 1978). La primera prioridad del Consejo de Europa era "was to provide a rallying point for the maintenance of pluralist parliamentary democracy and the protection of human rights" (Trim s.f.: 5). En la actualidad el Consejo de Europa cuenta con 47 países miembros (<http://www.coe.int>). España se

incorporó al Consejo de Europa el 24 de noviembre de 1977 (Instrumento de Adhesión de España … 1978).

La estructura y las principales instituciones del Consejo se constituyeron entre la fecha de su fundación y finales de los años 70. En concreto, las actividades de promoción de la diversidad lingüística y el aprendizaje de lenguas en el ámbito de la educación comenzaron a detallarse en el marco del tratado de la Convención Cultural Europea (European Cultural Convention), firmado en París en diciembre de 1954 y ratificado hasta la fecha por 49 estados. La relación completa de los tratados del Consejo de Europa puede consultarse en <http://conventions.coe.int/Treaty/Commun/ListeTraites.asp?CM=8&CL=ENG>. En su artículo 2, se recomienda a los estados signatarios que fomenten los estudios de idiomas, historia y civilización de los diferentes países que componen el Consejo.

Mientras que a finales de la década de los años sesenta Estados Unidos disponía, al menos, de una escala de niveles ampliamente extendida entre diversas instituciones para medir los niveles de competencia lingüística, ningún país europeo había desarrollado ninguna, aunque la principal preocupación del Consejo de Europa desde su fundación había sido la promoción, difusión y mejora de la enseñanza y de los métodos de idiomas. Estos fueron los temas centrales que se trataron en el Comité de Expertos Culturales (Committee of Cultural Experts) que se celebró entre el 4 y el 6 de noviembre de 1959 en París, así como de la Segunda Conferencia de Ministros de Educación reunidos entre el 12 y el 14 de abril de 1961 en Hamburgo. También, con el fin de cumplir con el objetivo de promocionar, difundir y mejorar de la enseñanza y de los métodos de idiomas, se crearon en 1962 el Consejo para la Cooperación Cultural (Council for Cultural Co-operation) y en 1964 el Centro de Documentación para la Educación (Documentation Centre for Education).

A pesar de todos estos esfuerzos seguía existiendo un desequilibrio y desde Europa se comenzó a mirar de reojo a los Estados Unidos. Bernard M. Schwarz, responsable de la educación permanente en la Universidad de Nancy, fue el primero que propuso como tema de debate el concepto de "unités capitalizables", es decir, la posibilidad de dividir los contenidos de una enseñanza global de la lengua en sus partes constituyentes, de manera análoga al sistema de *unit-credit* de los Estados Unidos (Ek 1975; Trim 1978, 1980). El Comité de Educación Extraescolar y Desarrollo Cultural del Consejo para la Cooperación Cultural (CDCC, en sus siglas en inglés) del Consejo de Europa celebró en 1971 en Rüschlikon (Suiza) un importante congreso titulado "The Linguistic content, means of evaluation and their interaction in the teaching and learning of modern languages in adult education" (Committee for Out-of-school Education and Cultural Development 1971) en el que se encomendó a un equipo de expertos sentar las bases para el desarrollo de un marco de referencia europeo para la enseñanza de lenguas a adultos e investigar acerca de viabilidad de implantar el sistema de *unit-credit* norteamericano en el ámbito europeo. Durante la celebración del congreso se debatieron cuestiones relacionadas con el sistema de unidades / crédito[4] como la forma de organizar los contenidos lingüísticos, los tipos de evaluación y los medios para implantar dicho sistema en la enseñanza/aprendizaje de las lenguas modernas a adultos. La posibilidad de implementar esta idea en el ámbito europeo despertó un amplio entusiasmo, aunque es preciso destacar que las ideas acerca de la estructura y los contenidos del sistema se encontraban en una fase inicial, motivo por el cual se decidió realizar un estudio de viabilidad y se creó un grupo de trabajo formado por John Trim, que ejercía la presidencia y que en aquel momento era el Director del Departamento de Lingüística de la Universidad de

[4] En el *Diccionario de términos clave de ELE* (bajo la entrada *Nivel Umbral*) se indica que el sistema de unidades/créditos consiste en determinar unas áreas de aprendizaje comunes y otras optativas en la enseñanza de idiomas a adultos europeos que no disponen de excesivo tiempo (entre 100 y 150 horas) y que únicamente desean alcanzar un nivel básico en la lengua que están aprendiendo.

Cambridge, René Richterich, Director del *Service de Recherche et d'Application* de Eurocentre, Jan van Ek, Director del Instituto de Lingüística Aplicada en la Universidad de Utrecht y David Wilkins, del Departamento de Lingüística de la Universidad de Reading, Reino Unido. Una de las primeras tareas del grupo consistió en encontrar la forma de dividir el concepto global de la lengua en unidades y subunidades a partir del análisis de alumnos adultos.

Como consecuencia de este impulso, entre 1972 y 1973 los integrantes de este equipo de expertos realizaron una serie de estudios fundamentales. Wilkins, por ejemplo, describe los contenidos lingüísticos y situacionales de la lengua inglesa tomando como base el sistema unidad / crédito y ofrece una primera clasificación en categorías nocionales y una división semántico-gramatical de las categorías de las funciones comunicativas; Van Ek menciona por vez primera el concepto *Threshold Level* en un sistema de unidades / crédito… Los estudios de los integrantes del equipo fueron publicados en 1973, aunque fueron reeditados posteriormente (Trim 1980 y Trim et al. 1980).

En una reunión posterior celebrada en Cambridge en enero de 1974 el grupo, reforzado por A. Peck y S. Hjelmstrom, prosiguió con los trabajos y redactó un documento que se distribuyó entre los comités de idiomas del Instituto de Cooperación Internacional de la Asociación Alemana de Educación de Adultos (Institut für Internationale Zusammenarbeit des Deutschen Volkshochschul-Verbandes, DVV en sus siglas en alemán) de la que Jan van Ek era asesor. Durante 1974 Van Ek profundizó en la definición de los requisitos lingüísticos mínimos que debe tener un usuario de la lengua para poder comunicarse en situaciones cotidianas con personas de otros países y en 1975 publicó *The Threshold Level in a European Unit/Credit System for Modern Lenguage Learning by Adults* (en adelante *Threshold Level*), en el que se define un determinado grado de dominio mínimo lingüístico.

El documento tuvo una gran repercusión tanto en el ámbito de la docencia como en el de la evaluación de lenguas en todo el ámbito europeo. Uno de los

conceptos en los que se basa *Threshold Level* es que independientemente de donde se encuentre un hablante de cualquiera de los estados miembros de la Unión Europea, e indistintamente de la lengua que este hable, sus necesidades de comunicación serán las mismas cuando se comunica con hablantes extranjeros en situaciones cotidianas. Por este motivo en Threshold Level se incluyen las situaciones en las que los hablantes pueden encontrarse cuando tratan de utilizar una lengua que no es la suya, los comportamientos lingüísticos que es posible que desarrollen y la capacidad lingüística exigida.

Ese mismo año, 1975, John Trim informó de que la traducción al español de Threshold Level la realizaría el profesor de español holandés Peter Jan Slagter. Con esta traducción se pretendía demostrar si era factible trasponer al español las descripciones que se habían formulado para el inglés sin que el modelo se alterara. El propio Peter Slagter lo relata en la introducción del siguiente modo:

> Se [me] pidió que elaborase una versión para el 'nivel umbral' español, tomando como base el modelo desarrollado por el grupo de expertos, y que, simultáneamente, investigase sobre las posibilidades de transposición a otra lengua de las descripciones elaboradas por el doctor Van Ek, cuyos ejemplos eran para el inglés. Esto quiere decir, en el fondo, que se me pidió dar una descripción de un 'nivel umbral' para el español siguiendo el modelo inglés lo más fielmente que mi experiencia como profesor y como autor de cursos y materiales didácticos para la enseñanza de lenguas me permitiese hacerlo. (Peter Slagter 1979: viii)

Sin embargo, antes de que el profesor Slagter finalizara la traducción, se publicó en Francia, en 1976, más que una traducción, una adaptación de *Threshold Level* al francés que llevaba por título *Un niveau seuil* (Coste, Courtillon, Ferenczi, Martins-Baltar y Papo 1976), en el que, además de las especificaciones de carácter lingüístico que presentaba el original en inglés, se incluían reflexiones sobre el aprendizaje y se proporcionaba un inventario mucho más rico de los exponentes que cubren una amplia gama de registros. La traducción al español de *Threshold Level,* más próxima a la versión

inglesa que a la adaptación francesa, la finalizó el profesor Slagter en 1979 con el título de *Un nivel umbral*. En los años 80 asistimos a una auténtica cascada de traducciones del documento: en 1982 al italiano, en 1983 al danés, en 1985 al holandés y en 1988 al noruego, al vasco y al portugués; en los 90 continuaron las traducciones con la que se hizo en 1993 al gallego, así como a otras muchas lenguas (galés, letón, estonio, lituano, maltés, ruso, griego, checo...). Doce años después de la primera edición, en 1991, P. J. Slagter, con la colaboración de J. L. M. Trim, publicó una nueva edición de la obra con el título *Threshold 1990*, versión que fue nuevamente revisada y corregida en 1998.

La repercusión que tuvo esta definición del nivel *Umbral* tanto en la didáctica como en la evaluación de lenguas extranjeras fue considerable. En el ámbito de la didáctica de ELE las instituciones educativas empezaron a diseñar programas nociofuncionales y comenzaron a publicarse métodos de enseñanza que sirvieran para realizar dichos programas (CVC, bajo la entrada de *Nivel Umbral*). En la evaluación de lenguas sirvió como modelo para la ideación de exámenes acreditativos del grado de competencia y dominio, como es el caso de los DELE o de los exámenes de la International Certificate Conference (ICC, en sus siglas en inglés), una organización internacional no gubernamental especializada en la enseñanza de lenguas extranjeras que establece estándares para una red internacional de estudiantes de idiomas. Sin embargo, la definición del nivel *Umbral* no estuvo exenta de críticas. Se le acusó de realizar una definición de nivel excesivamente baja, lo que a la larga, según afirmaban los críticos, podía provocar una disminución en el nivel alcanzado por los alumnos en un sistema de enseñanza que tenía como punto de referencia dicho nivel. De hecho este nivel, que actualmente equivaldría al B2, era el más bajo del sistema de unidades / crédito y fueron pocas las instituciones que en aquel momento reconocieron su validez.

Sin embargo, tras intentar implantar dicho nivel en las Universidades Populares de Viena (Viennese Volkshochschulen) se constató que, si bien era factible alcanzar el nivel *Umbral* tras un año de estudio en las destrezas receptivas, en las productivas únicamente lo conseguía un 33% del alumnado. Es decir, que la aplicación concreta de las especificaciones del nivel *Umbral* en un contexto educativo determinado, demostró que el problema o la crítica que podía hacerse al nivel *Umbral* era que resultaba excesivamente elevado para los estudiantes que acudían a las Universidades Populares de Viena y que muchos alumnos no conseguían alcanzar dicho nivel en un año de docencia regular. Se hizo necesario definir, en consecuencia, un nivel inferior al descrito por Van Ek (el tema del número de niveles que se podrían definir en el dominio de una lengua, desde el punto de vista teórico, ya había sido estudiado por el grupo formado tras el simposio celebrado en Rüschlikon en 1971). Por aquellas fechas se lanzó un curso para el aprendizaje del inglés titulado *Follow me*, coproducido por la BBC, Norddeutscher Rundfunk y Bayrischer Rundfunk y auspiciado por el Consejo de Europa. El curso quería ofrecer la posibilidad a los seguidores que lo finalizaran de obtener un certificado que acreditara el nivel de conocimientos lingüísticos alcanzado. Como el nivel *Umbral* resultaba excesivamente elevado, J. A. Van Ek y L. G. Alexander definieron en 1977 un nivel inmediatamente inferior al nivel *Umbral* al que denominaron *Waystage* (*Plataforma*, en la traducción al español). En 1991, Van Ek y Trim publicaron una nueva edición titulada *Waystage 1990*, versión que fue nuevamente revisada y corregida en 1998.

Tras la publicación de *Waystage* se hacía más evidente la carencia de una escala de niveles que pudiera servir de base para redactar nuevas definiciones de nivel. David Wilkins, uno de los integrantes del equipo, presentó en 1978 la descripción de una escala de siete niveles en la que se tenía en cuenta el grado de dominio que el hablante tenía en cada una de las cuatro destrezas o aptitudes lingüísticas. La denominación de los niveles era la siguiente:

Tabla 2
Escala de niveles de Wilkins, 1978

Nivel 7		Bilingüismo
Nivel 6		Competencia amplia
Nivel 5		Competencia media
Nivel 4		Competencia limitada
Nivel 3		Competencia básica (nivel *Umbral*)
Nivel 2		Competencia de supervivencia
Nivel 1		Aptitud para formular enunciados

Los trabajos para la definición de un nivel inmediatamente superior el nivel *Umbral* se demoraron hasta 1997, fecha en la que de nuevo Van Ek y Trim publicaron la definición del tercero de los niveles, al que denominaron *Vantage Level*, y que define un nivel inmediatamente superior al *Treshold Level.* Tras la definición de este nuevo nivel y con la reedición en 1991 de *Threshold 1990* y de *Waystage 1990*, se cierra el ciclo de la serie Treshold Level que se había iniciado en 1975.

Entre el 7 y el 14 de septiembre de 1977 se celebró en Ludwigshafen-am-Rhein (Alemania) una Conferencia Intergubernamental en la que el CDCC decidió cambiar de dirección y dar por finalizados los estudios que analizaban la posibilidad de implantar el sistema de unidades / crédito, lo que ocasionó que el pequeño grupo de expertos que se había constituido en el congreso de Rüschlikon fuera reemplazado por un equipo de 19 miembros que representaban a 13 países, también bajo la presidencia de John Trim.

El Consejo para la Cooperación Cultural del Consejo de Europa inició entre 1978 y 1981 el Proyecto Nº 4 titulado *Modern Language for the purpose of publicinsig and experimenting with the conceptual tools proposed by the experts*, cuyas actividades se expusieron en el documento *Modern Languages: 1971-1981* que se presentó en una conferencia celebrada en febrero de 1982. En el proyecto se indicaba el ámbito de aplicación de unos principios básicos en diferentes direcciones: "new methologies, new materials, multi-media systems, assessment and serf-assessment, learner autonomy, implications for language

teacher training" (Consejo de Europa 1981). Estas recomendaciones fueron consideradas de interés por el Consejo para la Cooperación Cultural y por el Comité de Ministros del Consejo de Europa, quien las incluyó en su recomendación *R(82)18* en la que se anima a los estados miembros a "adoptar o elaborar políticas nacionales en el campo del aprendizaje y la enseñanza de lenguas [con el fin de] conseguir una mayor convergencia a nivel europeo por medio de acuerdos adecuados para una continuada cooperación y coordinación de sus políticas", a fomentar "la colaboración nacional e internacional de instituciones gubernamentales y no gubernamentales que se dediquen al desarrollo de métodos de enseñanza y de evaluación en el campo del aprendizaje de lenguas modernas" y a tomar "las medidas necesarias para completar el establecimiento de un sistema eficaz de intercambio de información a nivel europeo que comprenda todos los aspectos del aprendizaje, la enseñanza y la investigación en el ámbito de las lenguas" (Consejo de Europa 1982).

La década de los años ochenta estuvo marcada por el *Proyecto 12 (1982-1987)* titulado *Learning and teaching modern languages for communication*, cuyo principal objetivo fue apoyar a los estados miembros a reformar la educación secundaria. Tras este proyecto, en 1990 se puso en marcha otro con el nombre: *Language learning for European citizenship*, en el que se quería dar prioridad a los sectores educativos que no habían recibido una atención preferente por parte de las autoridades europeas hasta aquel momento: Educación primaria, Bachillerato, Formación profesional y Educación de adultos. Al año siguiente, en 1991, se celebró en Rüschlikon, por iniciativa del Gobierno Federal Suizo, un simposio titulado: "Transparency and coherence in language learning in Europe: objectives, evaluation, certification" (Consejo de Europa 1992) en el que, por vez primera, se daban claras recomendaciones para el desarrollo de un documento que sirviera de marco de referencia en el ámbito europeo para el aprendizaje de la lengua en diversos niveles, el diseño de los cursos, la elaboración de métodos de enseñanza y la evaluación, con el

fin de promover y facilitar la cooperación entre los países de Europa, proporcionar los recursos necesarios para que pudiera realizarse con facilidad un reconocimiento mutuo de las certificaciones en Europa y ayudar a estudiantes, profesores, organizadores de cursos y administradores educativos. También se planteaba por vez primera la posibilidad de elaborar un portfolio europeo de lenguas (European Language Porfolio, ELP en sus siglas en inglés) en el que los alumnos pudieran registrar los cursos de lenguas realizados y sus experiencias en el aprendizaje de lenguas.

Para la redacción el *Marco de referencia* se creó un equipo de autores formado por JLM Trim (Director del proyecto), B. North y D. Coste, quienes presentaron en 1996 unos cuestionarios detallados que, tras ser analizados por unos 2000 expertos y recibir los comentarios de los gobiernos de los países miembros, de ONG y de la Unión Europea se modificaron sustancialmente. Esta revisión dio lugar a lo que se conoce como *Draft* 2. Esta segunda versión, junto con una propuesta de formatos de portfolio, se presentó en la Conferencia titulada *Language learning for European citizenship* (Trim 1997), en la que se pidió a la CDCC que revisara esta versión mediante la realización de aplicaciones piloto del *Marco de referencia.*

Tras realizar algunos cambios, en su mayoría de presentación, el Consejo de Europa (2001) publicó el proyecto en 2001, simultáneamente en inglés (con el título *Common European Framework of Reference for Languages: Learning, teaching, assessment)* y en francés, coincidiendo con el año europeo de las lenguas, entre cuyos objetivos figura la difusión del *Marco de referencia* y del PEL. En 2002 el Instituto Cervantes tradujo la obra al español con el título de *Marco Común Europeo de Referencia para las Lenguas: Aprendizaje, Enseñanza, Evaluación*[5]. La importancia que ha tenido y tiene el *Marco de referencia* en el ámbito de la certificación en general es de capital importancia. Se trata de un instrumento

[5] La traducción y adaptación que realizó el Instituto Cervantes se publicó en su página de internet en marzo de 2002.

esencialmente útil, centrado en la acción, que detalla de manera clara y precisa lo que los usuarios de la lengua son capaces de hacer como agentes sociales que realizan diversas tareas para lograr sus propósitos comunicativos.

La publicación del *Marco de referencia* ha generado una ingente documentación que resulta de gran ayuda tanto para el aprendizaje y la enseñanza como para la evaluación de las lenguas extranjeras. Se han publicado numerosos estudios en los que se glosa su contenido (García 2002) y se destaca y analiza la potencialidad de las directivas del *Marco de referencia* para la elaboración y el desarrollo de planes curriculares (Fernández 2003) para los redactores de métodos de enseñanza y para los profesores de idiomas (Llorián 2007).

En 2002 el Consejo de Europa publicó una guía para los usuarios del *Marco de referencia* (Trim et al. 2001). La guía está redactada con la intención de proporcionar ayuda a los profesionales que se acercarán al *Marco de referencia*: profesores, alumnos, formadores de profesores, autores de manuales, especialistas en evaluación, administradores educativos.

2.2.1. Escalas de niveles lingüísticos

En el ámbito europeo, tal y como señala Brian North (1997), fue a partir del último cuarto del siglo pasado cuando comenzaron a proliferar escalas de nivel que no estaban basadas en el modelo norteamericano. Numerosas instituciones, organismos y expertos comenzaron a dedicarse a la tarea de desarrollar escalas de dominio de la lengua. En el documento B6 del Anejo B del *Marco de referencia* (216) se enumeran las escalas de dominio que el proyecto suizo de investigación (ver abajo) utilizó como fuentes: escalas holísticas del dominio hablado general, para distintas actividades comunicativas, para las cuatro destrezas, de valoración para la evaluación oral y marcos de contenidos de programas y de criterios de evaluación para las fases pedagógicas del logro

de los objetivos (<http://www.oapee.es/oapee/inicio/ iniciativas/portfolio/ portfolio-esp.html>). Entre las escalas que se utilizan o se desarrollaron para el *Marco de referencia* destacamos las dos siguientes.

2.2.1.1. Escalas de Eurocentres (1983-1993)

Eurocentres es una institución establecida en Suiza especializada en la definición y la clasificación del dominio lingüístico por medio de escalas. Su Director Académico era ya por aquel entonces Brian North y para el desarrollo de las escalas de certificación de Eurocentres decidió centrarse en los usuarios con el fin de informar, de manera positiva, de lo que es capaz de hacer el alumno en cada una de las cuatro destrezas, y para asignar las descripciones de dominio de la lengua en los diferentes niveles partió directamente de los descriptores en lugar de hacerlo de las muestras de actuación. En una primera fase se realizó un análisis de lo que se deseaba describir y, posteriormente, se redactaron borradores de descriptores de las diferentes categorías. Este método resultó "especialmente adecuado para desarrollar descriptores de categorías relacionadas con el currículo, tales como las actividades lingüísticas comunicativas, pero también se puede utilizar para desarrollar descriptores relativos a la competencia" (Consejo de Europa 2002b: 200-201).

2.2.1.2. Proyecto suizo (1993-1996)

El proyecto suizo (Schneider y North 1999) fue financiado por el Consejo Nacional Suizo de Investigación Científica y se realizó como una continuación del congreso de Rüschlidon. El objetivo del proyecto era elaborar las escalas de descriptores que posteriormente se incluirían en los capítulos 3, 4 y 5 del *Marco de referencia* y diseñar los instrumentos de

autoevaluación precisos para el Portfolio Europeo de Lenguas. El proyecto se realizó en dos fases. La primera, que comenzó en 1994, se limitó al inglés como lengua extranjera y se centró en la evaluación realizada por el profesor, en la interacción y en la expresión. El segundo estudio, que comenzó en 1995, repitió parcialmente el estudio de 1994, aunque incorporó la comprensión, se añadieron el dominio del francés y del alemán y la evaluación que realiza el profesor se sumó la autoevaluación de los estudiantes y la información sobre exámenes (Cambridge —inglés—, DELF/DALF —francés— y Goethe —alemán—). (Consejo de Europa 2002b: 209).

El proyecto se desarrolló en cuatro etapas: una fase intuitiva que consistió en analizar y deconstruir diversas escalas de domino existentes en el ámbito europeo; una fase cualitativa en la que se analizó el trabajo de un equipo de profesores que estudiaban las actuaciones en vídeo; una fase cuantitativa en la que profesores evaluaban a los alumnos mediante una serie de cuestionarios (siete el primer año y cinco el segundo) y, finalmente, una fase interpretativa en la que se realizaron diversos trabajos con las escalas de descriptores. Afirma North (1997) que el objetivo principal del proyecto era utilizar un sistema de trabajo mediante un banco de ítems con el fin de redactar un banco de descriptores con información de sus coeficientes de dificultad. Este banco de descriptores fue utilizado para la primera edición de los niveles comunes de referencia del *Marco de referencia* en 1996 (Consejo de Europa 1996a) y para la creación de los instrumentos de auto-evaluación de un prototipo de Pasaporte o Portfolio de lenguas (Consejo de Europa 1997).

Los trabajos en relación con las escalas aún continúan. En 2015 el Consejo de Europa ha auspiciado el proyecto titulado "CEFR descriptors for Mediation", financiado por Eurocentres y liderado por Brian North, sobre la mediación lingüística. El proyecto tiene como objetivo crear descriptores lingüísticos de mediación y adaptarlos-introducirlos en las escalas del *Marco de referencia*.

2.2.1.3. Escalas descriptivas de niveles lingüísticos desarrolladas independientemente del *Marco de referencia*

Para llegar al primer borrador del *Marco de referencia* se había recorrido, como se ha indicado arriba, un largo camino. Antes de disponer de esta versión, varias universidades e instituciones europeas habían comenzado a trabajar en proyectos con la intención de redactar una serie de especificaciones que sirvieran para definir los diferentes niveles del aprendizaje o del dominio de las lenguas. Los dos proyectos más relevantes son: Can-Do de ALTE y DIALANG. Ambos, tal y como reconoce el *Marco de referencia* a lo largo de sus páginas, están relacionados y utilizan una metodología similar en su desarrollo. En los dos casos, la fuerza arrolladora con la que surgió el *Marco de referencia*, provocó que en la fase final de ambos proyectos se ampliaran y adaptaran los descriptores del *Marco de referencia* y se hiciera una aplicación práctica de los niveles comunes de referencia.

2.2.1.3.1. El proyecto Can Do statements de ALTE (1992-2002)

Desde sus orígenes en 1989, uno de los principales objetivos de ALTE ha sido establecer escalas de niveles comunes de competencia con el fin de facilitar el reconocimiento transnacional de las diversas certificaciones otorgadas por las diferentes instituciones certificadoras europeas. El resultado de este proyecto son las especificaciones *Can Do statements* ('Puede hacer'), un proyecto de investigación a largo plazo que se inició en 1992 y que ha recibido financiación por parte de la Unión Europea. El objetivo del proyecto "es desarrollar y validar un conjunto de escalas relacionadas con la actuación donde se describa lo que los alumnos son realmente capaces de hacer en la lengua extranjera" (Consejo de Europa 2002b: 236).

Para la realización de las escalas se partió, primeramente, de una descripción de los usuarios y se determinaron sus principales intereses; seguidamente, se

utilizaron las descripciones de los niveles publicadas por el Consejo de Europa (*Umbral* y *Plataforma*) para redactar las especificaciones iniciales y se reajustaron estas según su adecuación con los candidatos y finalmente se experimentaron con alumnos y profesores y se realizaron las oportunas revisiones y simplificaciones (ALTE, 2002).

En el momento de comenzar a redactar el proyecto se definieron cinco niveles (aún no estaba estandarizada la escala de seis niveles del *Marco de referencia*;) mediante unas cuatrocientas especificaciones organizadas en tres áreas generales: Sociedad y turismo, Trabajo y Estudio (áreas que más interesan a la mayoría de los alumnos de idiomas) y cada una de las áreas incluye otras secciones más concretas. Todas ellas presentan hasta tres escalas para las siguientes destrezas:

- Comprensión Auditiva/Expresión Oral.
- Comprensión de Lectura.
- Expresión escrita.

Estos cinco niveles se relacionaron con los del *Marco de referencia* una vez que estos se hicieron públicos del siguiente modo:

Tabla 3
Escala de niveles de MCER/ALTE

MCER	A1	A2	B1	B2	C1	C2
ALTE	ALTE Breaktrough Level	ALTE Level 1	ALTE Level 2	ALTE Level 3	ALTE Level 4	ALTE Level 5

MCER: *Marco de referencia*
Tabla basada en ALTE (2007)

Estas escalas son multilingües y han sido traducidas, hasta ahora, a trece de las lenguas representadas en ALTE: alemán, catalán, danés, español, finés, francés, holandés, inglés, griego, italiano, noruego, portugués y sueco. Su principal objetivo es constituir un marco de referencia por medio del cual se puedan comparar y relacionar exámenes de distintos niveles y se puedan

demostrar las equivalencias existentes entre los sistemas de certificación de los miembros de ALTE en términos referidos a destrezas lingüísticas de la vida cotidiana que podrán realizar los candidatos que se presenten a los exámenes de dominio (Consejo de Europa 2002b: 236). Neil Jones demostró en 2002 la utilidad de las vinculaciones de los Can-Do con el *Marco de referencia*.

2.2.1.3.2. DIALANG (1996-2002)

El proyecto DIALANG se llevó a cabo con la ayuda financiera de la Comisión Europea y la Dirección General de Educación y Cultura (Programa SOCRATES, LINGUA, Acción D) y en él colaboraron diferentes universidades europeas, por ejemplo el Centre for Applied Language Estudies de la Universidad de Jyväskyla (Finlandia), la Freie Universität de Berlín (Alemania) y la Universidad de Lancaster (Reino Unido). Se desarrolló en dos fases entre los años 1996 y 2002 y comenzó a funcionar en octubre de 2006. Desde el año 2009, cuando el sitio web DIALANG fue atacado por piratas informáticos, el sistema está alojado en la Universidad de Lancaster. La dirección electrónica actual para acceder al proyecto es: <http://www.lancs.ac.uk/researchenterprise/dialang/about>

El equipo de trabajo de DIALANG (formado por Alex Teasdale —presidente—, Neus Figueras, Ari Huhta, Fellyanka Kaftandjieva, Mats Oscarson y Sauli Takala) extrajo la mayor parte de las especificaciones de autoevaluación del *Marco de referencia* (Versión 2, inglesa) (Consejo de Europa 1996a) y seleccionó de la versión de 1998 los enunciados más "concretos, claros y sencillos" (Consejo de Europa 2002b: 218-219). También examinó la tesis doctoral de North (1996/2000) con el fin de seleccionar las especificaciones de las cuatro destrezas, aunque las correspondientes a la expresión oral —destreza que no forma parte del sistema— no se validaron.

Huhta et al. (2002) describen el procedimiento utilizado por el equipo de DIALANG para utilizar el *Marco de referencia.*

DIALANG está diseñado para que cualquier usuario pueda realizar una evaluación de diagnóstico de las siguientes lenguas: alemán, danés, español, finés, francés, griego, holandés, inglés, irlandés, islandés, italiano, noruego, portugués y sueco. Además de determinar el nivel lingüístico, el usuario de DIALANG debe realizar una autoevaluación y, tras realizar las pruebas lingüísticas, el sistema le ofrecerá una retroalimentación en forma de consejos acerca de cómo mejorar sus destrezas lingüísticas (Puig 2008). El sistema no emite ninguna certificación tras su utilización.

El usuario, tras entrar en el sistema, elige la lengua en la que quiere recibir las instrucciones y los consejos finales. Seguidamente debe realizar una prueba de clasificación y seleccionar la destreza en la que quiere evaluarse. Después debe autoevaluarse, para lo cual tiene que seleccionar una serie de especificaciones indicando si considera que puede realizarlas o no y finalmente realizará la prueba. En la fase de retroalimentación se preguntará al usuario si el nivel que obtuvo en la autoevaluación se corresponde con el obtenido al realizar la prueba.

La duración media de una prueba completa, incluyendo el test de clasificación, prueba de idioma y las tareas de autoevaluación es de 30-45 minutos; si únicamente se realiza la prueba de idioma la duración típica es de 20-30 minutos. Cada prueba consta de alrededor de 30 ítems y no hay tiempo máximo para marcar las respuestas, aunque las audiciones únicamente se pueden escuchar una vez. Según el resultado de la prueba de nivel, el sistema decidirá si presenta una versión fácil, intermedia o difícil; si no se realiza la prueba de nivel el sistema selecciona por defecto la versión intermedia. Al finalizar la prueba, DIALANG informa del nivel (siguiendo la escala de niveles del *Marco de referencia*), de las respuestas correctas e incorrectas (se pueden volver a ver las preguntas) y ofrece consejos en forma de cuadros

comparativos con el nivel inferior y superior al obtenido en relación con el aspecto evaluado.

2.2.2. El Portfolio Europeo de Lenguas (PEL)

En la primavera de 1996, al finalizar el proyecto patrocinado por el Consejo Nacional Suizo de Investigación Científica (Proyecto suizo), se publicó un primer borrador completo de un portfolio europeo de lenguas (Schneider, G. y North, B., 1999 y 2000) y en 1997, el Consejo de Europa, tras la publicación del segundo borrador del *Marco referencia*, analizó la posibilidad de utilizar el portfolio europeo de lenguas en la enseñanza de idiomas (Consejo de Europa, 1997). Entre 1997 y 2000 (Ministerio de Educación, Cultura y Deporte, 2003) se pilotaron diversos proyectos piloto y en octubre de 2000 la Conferencia Permanente de los Ministros de Educación de todos los estados miembros del Consejo de Europa adoptó una recomendación en la que quedaban establecidos los principios y directrices del portfolio. A medida que finalizaba la fase piloto se elaboraron dos documentos de referencia: las directrices para diseñar e implementar el portfolio (Consejo para la Cooperación Cultural, 2000a) y el reglamento para acreditar los modelos del portfolio (Consejo para la Cooperación Cultural, 2000). También se redactaron las guías para los redactores del portfolio (Schneider y Lenz, 2001) y para los profesores y formadores de profesores (Ministerio de Educación, Cultura y Deporte, 2003). En el año 2000 se constituyó un Comité de Validación compuesto por nueve miembros con el encargo de que las diferentes versiones del PEL cumplieran con las directrices del Consejo de Europa. Desde 2001 hasta 2010 se validaron 118 portfolios para todas las fases educativas: primaria, secundaria, formación profesional, adultos…. La experiencia acumulada por el Consejo de Europa a lo largo de estos años le permitió diseñar un conjunto de plantillas y recursos para que cualquier interesado, mediante una auto declaración, pueda registrar su modelo (<http://www.coe.int/t/dg4/education/elp/ELP-REG/Default_EN.asp>). No obstante, a pesar de estas cifras, el número medio de copias en uso por cada uno de los portfolios

validados es de 6.600, lo que evidencia las dificultades existentes entre las autoridades educativas para lograr un uso generalizado y extendido (Little, Goullier y Hughes 2011:5).

El PEL consiste en un documento personal en el que el alumno puede llevar un registro de su aprendizaje de lenguas, así como de los diferentes diplomas, certificados o títulos que haya conseguido durante el aprendizaje de la lengua. Consta de 3 partes:

- Pasaporte de Lenguas. El titular del pasaporte debe reflejar lo que sabe hacer en las lenguas que aprende o ha aprendido. Para ello debe reflexionar y autoevaluarse por medio de un cuadro de autoevaluación en el que se describen las competencias por destrezas. También debe reseñar los contactos que haya tenido con otras lenguas y culturas, los cursos realizados y los certificados que ha obtenido.
- Biografía lingüística. Se trata del apartado en el que el alumno debe describir sus experiencias en cada una de las lenguas que aprende / ha aprendido. Sirve para que el alumno pueda planificar y evaluar su progreso adecuadamente.
- *Dossier*. Sirve para que el alumno adjunte los certificados y diplomas que ha obtenido a lo largo de su proceso de aprendizaje, así como los trabajos personales (trabajos escritos, proyectos, grabaciones en audio y/o vídeo, presentaciones, etc.) que ha realizado.

El PEL tiene como principal objetivo animar a los alumnos en su proceso de aprendizaje de lenguas, facilitar su movilidad en el ámbito europeo y promover el entendimiento y la tolerancia entre los ciudadanos de Europa.

En España, el proyecto está liderado por la Agencia de Programas Educativos Europeos. En marzo de 2001 se constituyó en España un Comité Nacional para el desarrollo de un Portfolio Europeo de las Lenguas (PEL) que encargó a un equipo de expertos en enseñanza de idiomas la elaboración y el diseño de un Portfolio europeo de las lenguas para aplicarlo en España. En febrero de 2003 este equipo presentó cuatro prototipos de PEL para cuatro grupos de

edades y niveles educativos distintos a representantes de las Comunidades Autónomas. Tras incorporar las aportaciones realizadas, se remitieron en junio de 2003 las propuestas al Comité de Validación Europeo del Consejo de Europa, órgano que finalmente validó el 7 de noviembre de 2003 tres de los modelos presentados: el PEL para alumnos de 3 a 7 años, el PEL para alumnos de 8 a 12 años y el PEL para Enseñanza Secundaria, F.P. y Bachillerato (12-18 años) y durante 2004 validó también el cuarto de los PEL, el de Adultos.

2.2.3. *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*

La aparición del *Marco de referencia* puso en evidencia que había una falta de rigor tanto en la planificación u organización de cursos de enseñanza de idiomas como en el diseño de métodos de enseñanza y en la definición de niveles de los exámenes de dominio. En 2001, el mismo año de la aparición del *Marco de referencia*, se celebró un Congreso Internacional de ALTE en Barcelona. En una de las sesiones, el profesor de los Países Bajos, John H.A.L. de Jong, del Servicio de Exámenes de Idiomas (Language Testing Services) (Jong 2001) cuestionó los procedimientos utilizados por algunas de las instituciones que certifican niveles lingüísticos en Europa para validar sus respectivos exámenes. Estas críticas calaron entre la comunidad de especialistas y en julio de 2002 el Ministerio de Educación finlandés organizó un seminario en Helsinki en el que se incrementaron los ataques lo que ocasionó que el Consejo de Europa pusiera en marcha un equipo de trabajo (formado por Brian North —director—, Neus Figueras, Sauli Takala, Piet van Avermaet, ALTE y Norman Verhelst.) al que encargó la redacción de un manual para vincular exámenes con la escala de niveles del *Marco de referencia*

(Figueras 2008). En 2003 el Consejo de Europa publicó la edición piloto del *Manual* y su aparición puso de manifiesto las dificultades de instituciones y redactores y organizadores de exámenes de dominio utilizar el *Marco de referencia* como en la definición de los niveles de sus exámenes. En 2007, en el Foro Intergubernamental que convocó el Consejo de Europa titulado *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities*, se constató que, tras la presentación de los resultados de una encuesta realizada entre mayo y septiembre de 2006 con la que se pretendía obtener información acerca del uso que se hacía del *Marco de referencia* en los estados miembros del Consejo de Europa (Martyniuk y Noijons 2007), la problemática alcanzaba a todo el ámbito de la Unión Europea. En el foro también se analizaron algunos de los problemas que el *Marco de referencia* presentaba para su utilización: la ausencia de dogmatismos del *Marco de referencia*, así como su naturaleza abierta, flexible y no específica, había provocado cierta perplejidad entre sus usuarios y la utilización prioritaria de la dimensión vertical del *Marco de referencia* en detrimento de la horizontal había provocado usos excesivamente simplistas y superficiales (Figueras 2008). En 2009, tras la contribución de diversos especialistas e instituciones, en su mayoría de Europa, se publicó la versión definitiva del *Manual* (Consejo de Europa 2009a).

El principal objetivo del *Manual* es lograr que las entidades certificadoras que redactan exámenes de dominio en el ámbito europeo apliquen procedimientos transparentes y prácticos para vincular sus exámenes con el *Marco de referencia* con el fin de incrementar la transparencia en los contenidos de los exámenes. El *Manual* se divide en cuatro grandes apartados (Figueras 2004):

- Familiarización: presenta actividades para asegurar que los participantes en los procesos de análisis o anclaje de las pruebas con el *Marco de referencia* conocen en profundidad los contenidos y niveles que describe el *Marco de referencia*.

- Especificación: se refiere a la relación entre los contenidos del examen que se quiere vincular y las categorías que presenta el *Marco de referencia*.
- Estandarización: incluye procedimientos estandarizados que aseguran una interpretación común de los *Niveles comunes de referencia* con ejemplos concretos de tareas de examen y de actuaciones de alumnos con el fin de garantizar la solidez de los juicios que se adopten al calificar o redactar las tareas.
- Validación empírica: se refiere a la recogida de datos y el análisis de resultados con el fin de poder demostrar la solidez tanto del examen como del anclaje de este con el *Marco de referencia*.

El *Manual* está acompañado por un suplemento (Consejo de Europa 2009b) que tiene como principal objetivo proporcionar a los usuarios del *Manual* información adicional para vincular sus exámenes con el *Marco de referencia*. Consta de nueve secciones en las que se tratan temas como la Teoría Clásica de los Tests, la teoría de la Generalizabilidad, la Teoría de Respuesta al Ítem (TRI) o MFRM entre otros.

Además del *Manual*, el Consejo de Europa ha promovido diverso material complementario con el fin de proporcionar diversas herramientas que coadyuven en el objetivo de mejorar la calidad de los sistemas de certificación en el ámbito europeo:

- Fichas o formularios detallados para el estudio de cada una de las actividades lingüísticas de comunicación (capítulo 4) del *Marco de referencia* y para los aspectos relacionados con las competencias comunicativas (capítulo 5) de los exámenes o de las pruebas que se quieren analizar.
- Material adicional sobre el mantenimiento de los estándares de las lenguas, contextos y administraciones mediante el juicio de profesores y las escalas TRI (North y Jones 2009).

2.3. Otros proyectos

2.3.1. Listas de control para el análisis de los contenidos

Para caracterizar el contenido tanto de las pruebas como de las tareas que conforman las pruebas de expresión oral, expresión escrita, comprensión auditiva y comprensión de lectura, ALTE ha puesto a disposición del Consejo de Europa las listas de control (*Checklists*) que han desarrollado sus miembros. Estas listas de control permiten chequear por medio de procedimientos estandarizados las características que poseen las pruebas y tareas con el fin de poder extraer información fácilmente comparable entre exámenes de una misma institución y de instituciones diferentes. El Consejo de Europa ha seleccionado las listas de control de ALTE para analizar las destrezas productivas: expresión oral y escrita (ALTE members 2005a, 2005b, 2007a y 2007b). ALTE también había desarrollado con anterioridad listas de control para el análisis tanto de la expresión oral y escrita como de la comprensión de lectura y auditiva y de la competencia estructural de las diferentes versiones de las pruebas y de cada una de las tareas que componen dichas pruebas (ALTE 2009a, 2009b, 2009c, 2009d, 2009e, 2009f, 2009g, 2009h, 2009i, 2009j y 2009k), mientras que para analizar la comprensión de lectura y auditiva ha seleccionado las elaboradas por el Dutch CEFR Construct Group Project. Los responsables del proyecto (coordinado por Charles Alderson, de la Universidad de Lancaster y redactado por Neus Figueras —Departamento de Educación de la Generalitat de Cataluña—, Henk Kuijper —CITO, Países Bajos—, Claire Tardieu, —IUFM, París, Ministerio de Educación Nacional de Francia—, Guenter Nold, —Deutsche Institut für Internationale Pädagogische Forschung, de la Universidad de Dortmund, Sauli Takala, ex miembro de la Universidad de Jyväskylä—) (Alderson, Figueras, Kuijper, Tardieu, Nold y Takala 2006) habían recibido "el encargo del gobierno holandés de estudiar los contenidos del *Marco de referencia* y facilitar la elaboración de especificaciones para pruebas de inglés, francés y alemán"

(Figueras 2008) con el objetivo de desarrollar un instrumento que ayudara a los responsables de exámenes y redactores de tareas a relacionar las tareas de comprensión de lectura y auditiva con el *Marco de referencia*. Por medio de una lista de control con elementos desplegables se detallan las características de los textos y de los ítems. El proyecto, que incluye algunos elementos que no figuran en el *Marco de referencia* es, en muchos aspectos, complementario al *Manual,* y permite a los redactores de exámenes analizar las pruebas de comprensión de lectura y auditiva con el fin de relacionarlas con el *Marco de referencia*.

2.3.2. Ejemplos de producciones orales

Con el fin de garantizar que los niveles que describe el *Marco de referencia* se interpretan de igual forma por los profesionales de la lengua, la División de Políticas Lingüísticas del Consejo de Europa celebró un seminario, organizado por el Centro Internacional de Estudios Pedagógicos de Francia (Centre International d'Études Pédagogiques, CIEP en sus siglas en francés), el 23-25 de junio de 2008 fruto del cual se desarrolló un material en DVD con ejemplos de producciones orales en alemán, inglés, francés, español e italiano con los que se ilustran cada uno de los seis niveles del *Marco de referencia* y que cuentan con el consenso de un amplio grupo de evaluadores. Toda la información relativa al seminario, así como a los análisis realizados puede consultarse en: <http://www.coe.int/t/DG4/Portfolio/?L=E&M=/main_pages/illustrationsse.html>. (cf. Breton, Lepage y North 2008). Para realizar los análisis de las pruebas se utilizaron las listas de control para el análisis de contenidos desarrolladas por ALTE, aunque también se tuvieron en cuenta las del Dutch CEFR Construct Group Project.

2.3.3. Niveles de referencia para el español

Los redactores del *Marco de referencia* son conscientes de que en el *Marco de referencia* no se presentan descripciones detalladas de microfunciones, formas gramaticales y vocabulario (Consejo de Europa 2002b: 28) y remiten a las especificaciones que se han realizado para lenguas concretas (Ek y Trim 1991). La serie Threshold Level, como ya hemos indicado, únicamente había descrito 3 niveles; es decir, faltaban los otros tres. En el ámbito hispánico, la institución de referencia que tiene encomendada por ley la promoción universal de la enseñanza y el uso del español, el Instituto Cervantes, tenía pendiente la renovación de su Plan Curricular que había publicado en 1994 (Instituto Cervantes, 1994). Estas dos coincidencias, además de los impulsos de las conferencias auspiciadas por el Consejo de Europa, especialmente "Language Policies for a Multilingual and Multicultural Europe 1997-2000" (<http://www.coe.int/t/dg4/linguistic/ historique_en.asp>), ocasionaron que a los dos años de la publicación de la traducción al español del *Marco de referencia*, en 2004, el Instituto Cervantes comenzara los trabajos para la redacción de los Niveles de referencia para el español que el Instituto ha incorporados como propios en la actualización de su Plan curricular con el fin de actualizar el antiguo que estaba basado, esencialmente, en el planteamiento teórico del nivel *Umbral* (puede consultarse la historia de la redacción del proyecto en Instituto Cervantes 2007: 47-51). Para su desarrollo el Instituto Cervantes tuvo en cuenta las directrices del Departamento de Política Lingüística del Consejo de Europa (2005) que se plasmaron en una guía que sirviera de referencia a las instituciones que realizaban las descripciones para sus respectivas lenguas.

Esta situación comenzó a cambiar en 2001 cuando el Consejo de Europa publicó el *Common European Framework for Languages: Learning, Teaching, Assessment*. Seis años más tarde, en 2007, el Instituto Cervantes publicó el *Plan curricular del Instituto Cervantes. Niveles de referencia para el español* (en adelante *Plan curricular*), obra en la que se fijan y desarrollan los distintos niveles de dominio

que pueden establecerse en un programa de enseñanza del español como lengua extranjera según la escala de niveles que propone el *Marco de referencia*.

A raíz de la publicación del *Marco de referencia* en 2001, el Instituto Cervantes recoge la petición que en este se hacía de que "se desarrolle, para cada lengua nacional y regional europea el material lingüístico necesario en relación con los descriptores que caracterizan las competencias comunicativas de los alumnos en los diferentes niveles así como las especificaciones que correspondan a las competencias generales". (Instituto Cervantes 2007: 13) y elaboró su *Plan curricular*.

En otoño de 2004 (*Plan curricular*: 48) el Instituto Cervantes comenzó a trabajar en la actualización de su Plan curricular con la colaboración de los equipos docentes de los diversos centros del Instituto Cervantes y el asesoramiento de profesores y expertos en el ámbito de la enseñanza del español. En una primera fase se elaboraron borradores de las listas de Gramática, de Objetivos generales y de los inventarios de Funciones y Nociones generales. Para su elaboración se tomó como base principal el *Marco de referencia* y los tres repertorios de descripciones de los niveles de competencia lingüística de la serie del nivel *Umbral*. Tras una reunión celebrada en el mes de diciembre de ese mismo año entre el Departamento de ordenación académica y los responsables académicos de los equipos docentes de los centros del Instituto Cervantes se revisaron diferentes aspectos de los borradores de inventarios y se continuó con la redacción de nuevos inventarios entre un equipo de profesores del propio Instituto y el Departamento de ordenación académica. En 2005 se celebró en Estrasburgo un seminario de trabajo en el que se reunieron representantes de instituciones de las lenguas nacionales y regionales de Europa en el que se sentaron las bases sobre las que se debería basar el desarrollo de los Niveles de referencia (Consejo de Europa 2005). En diciembre de 2005 el Departamento de Política Lingüística del Consejo de Europa, a partir de las conclusiones de

seminario publicó la *Guía para la elaboración de descripciones de niveles de referencia para las lenguas nacionales y regionales* (Reference Level Descriptions for national and regional languages, RLD en las siglas en inglés) (Consejo de Europa 2005) con el fin de garantizar la transparencia y coherencia de los desarrollos de descripciones de niveles de referencia que estaban realizando o iban a realizar diferentes instituciones de países europeos. Casi de forma paralela, en otoño de 2005, había dado comienzo la segunda fase de los trabajos coordinados por el Instituto Cervantes en la que se revisó y depuró el material de los primeros borradores. Posteriormente, tras fijarse los segundos borradores, un equipo de profesores externos de prestigio revisó el conjunto del material. Finalmente, a partir de la primavera de 2006, el Consejo de Redacción de la elaboración del *Plan curricular* comenzó la elaboración de las versiones definitivas de los diversos borradores. Puede consultarse la historia pormenorizada de la elaboración del *Plan curricular* en la Descripción del desarrollo del proyecto del *Plan curricular* (pp. 47-51).

2.3.4. *Manual for Language Test Development and Examining*

El *Manual for Language Test Development and Examining* (Consejo de Europa 2011) (en adelante *Manual for Language Test*) es una versión revisada de un documento publicado en 2002 (Consejo de Europa 2002a) por el Consejo de Europa titulado *Language examining and test development* (2002a) que, a su vez, era una revisión de otro elaborado en 1996 y titulado *Users' Guide for Examiners* (Consejo de Europa 1996b), que fue una de las guías encargadas por el Consejo de Europa para acompañar al primer borrador del *Marco de referencia* (Consejo de Europa 1996a). El equipo de redacción de esta revisión estaba formado por David Corkill, Michael Corrigan, Neil Jones, Michael Milanovic, Martin Nuttall y Nich Saville. Muchos de los miembros de ALTE han colaborado en el proceso.

El *Manual for Language Test* tiene como principal objetivo complementar el *Manual* y se centra en aspectos relacionados con el desarrollo de las pruebas de examen que este no trata. Su principal finalidad es proporcionar una guía coherente para el desarrollo de un examen en la que presenta el diseño de pruebas de examen como un proceso cíclico en el que cada etapa está relacionada con el trabajo realizado en la anterior (p.18). En las diferentes secciones en las que se divide el *Manual for Language Test* se tratan los siguientes aspectos:

- Definición de conceptos como validez, fiabilidad, aspectos éticos.
- Análisis del desarrollo de los exámenes.
- Construcción de los exámenes.
- Administración de los exámenes.
- Presentación de los resultados.
- Seguimiento y revisión del proceso cíclico con el propósito de mejorar la calidad y la utilidad del examen.

C A P Í T U L O 3

Utilización del *Marco de referencia* y del *Plan curricular* para el diseño y elaboración de pruebas de nivel

3.1. Enfoque adoptado

El *Marco de referencia* destaca que en el proceso de la comunicación se realizan tareas que requieren del uso de estrategias. El enfoque que describe el *Marco de referencia* está claramente centrado en la acción y se concreta "en la relación existente entre […] el uso que los agentes hacen de las estrategias ligadas a sus competencias y la manera en que perciben o imaginan la situación y […] la tarea o las tareas que hay que realizar en un contexto específico bajo condiciones concretas" (Consejo de Europa 2002b: 15). Lo que se pretende en una prueba de dominio es realizar inferencias en relación con el grado de desarrollo de las competencias comunicativas de la lengua del candidato y de su capacidad de ponerlas en funcionamiento en tareas que pretenden simular las situaciones que se producen en la *vida real*. Para lograr este propósito es preciso recoger muestras de actuación de los candidatos mediante la realización de tareas estandarizadas que sean representativas de las que en la *vida real* sean significativas para los candidatos que se presentan al examen de dominio.

Para que este proceso se realice de forma coherente, es necesario, en consecuencia, describir y detallar en las especificaciones de la prueba todos los aspectos susceptibles de ser evaluados así como el grado de dominio que tendrán que demostrar en relación con ellos los candidatos. En concreto, será preciso explicitar los diversos elementos de las competencias que los

candidatos deberán activar al realizar las actividades comunicativas de la lengua al realizar tareas. Para describirlos habitualmente se parte de la taxonomía del Marco de referencia y del desarrollo que el Instituto Cervantes ha hecho en su *Plan curricular*.

3.2. Modelo de competencia comunicativa

El concepto de competencia comunicativa fue acuñado por el etnolingüista Dell Hymes en 1971, aunque es más conocido por el artículo publicado al año siguiente en la obra colectiva de Pride y Holmes (Hymes 1972; existe traducción al español del artículo en Llobera 1995). Hymes propone que la competencia comunicativa debe entenderse como un conjunto de habilidades y conocimientos que permite que los hablantes de una lengua se entiendan. Este concepto fue ampliado y perfeccionado algunos años más tarde, primero por M. Canale y D. Swain (1980, 1981 y 1983) —quienes incluyen cuatro áreas de conocimiento y habilidad (o cuatro dimensiones) dentro de la competencia comunicativa: competencia gramatical, competencia sociolingüística, competencia discursiva y competencia estratégica—, y posteriormente por Bachman en 1990 y por Bachman y Palmer en 1996 (Gutiérrez 2008).

El modelo de competencia comunicativa amplió considerablemente las perspectivas de los evaluadores en los años 80, ya que proporcionaba un marco para la descripción y para la validación que anteriormente no estaba disponible. J. Van Ek sigue en 1986 el modelo de Canale y Swain (1981), aunque incrementa el número de competencias al incluir la competencia sociocultural y la competencia social.

En 1990: 81-110, L. F. Bachman lanza su propuesta de "habilidad lingüística comunicativa", que se basa claramente en la propuesta desarrollada por

Canale y Swain, en la que se definen dos grandes áreas de competencia comunicativa:

- Competencia organizativa:
  - Competencia gramatical: vocabulario, morfología, sintaxis, fonología/grafología.
  - Competencia textual: cohesión y organización retórica.
- Competencia pragmática:
  - Competencia ilocutiva: funciones ideativas, manipulativas, heurísticas, imaginativas.
  - Competencia sociolingüística: sensibilidad hacia las diferencias de dialecto o de variedad, sensibilidad hacia las diferencias de registro, sensibilidad a la naturalidad, habilidad para interpretar referencias culturales y lenguaje figurado.

El conocimiento estratégico en el modelo de L. F. Bachman actúa como un componente externo a lo que él denomina competencia lingüística (denominada competencia comunicativa por la mayor parte de autores), es decir, que no forma parte de ninguna de las competencias (ni de la organizativa ni de la pragmática) y se refiere directamente al conocimiento procedimental. Para Bachman, este conocimiento estratégico se refiere a una habilidad que permite utilizar el resto de habilidades de la manera más efectiva posible al realizar una determinada tarea.

En 1996, Bachman y Palmer presentan una versión modificada de este modelo de competencia lingüística en el que se señala que aunque el conocimiento lingüístico es un requisito previo esencial para la utilización del lenguaje, su aplicación correcta se ve afectada principalmente por la interacción con otros procesos mentales, como esquemas de conocimiento y esquemas afectivos. Los primeros, los esquemas de conocimiento, corresponden a los conocimientos y a la experiencia del mundo, mientras que

los segundos, los afectivos, se refieren a los recuerdos emocionales. El modelo indica que estas tres características mentales se activan y se utilizan por medio de un proceso metacognitivo que implica a las tres fases: estrategias de evaluación, estrategias de planificación y estrategias de fijación de objetivos[6].

El *Marco de referencia*, aunque amplía y enriquece toda esta taxonomía de competencias, deja de concebir el componente estratégico como uno de los integrantes de la competencia comunicativa y lo define como lo que pone "en funcionamiento las actividades y las competencias mentales en el curso de la comunicación" (Consejo de Europa 2002b: 90). La utilización de las estrategias, para el *Marco de referencia*, consiste en la aplicación de los principios cognitivos de planificación, ejecución, control y reparación de los diversos tipos de actividad comunicativa: expresión, comprensión, interacción y mediación (Consejo de Europa 2002b: 61). En el *Marco de referencia* (Consejo de Europa 2002b*:* 13-14, 106 y ss.) la competencia comunicativa está integrada por los siguientes componentes: competencia lingüística, competencia sociolingüística y competencia pragmática.

La competencia lingüística se subdivide, a su vez, en competencia léxica, gramatical, semántica, fonológica, ortográfica y ortoépica. La sociolingüística comprende el conocimiento y las destrezas necesarias para abordar la dimensión social del uso de la lengua y está compuesta por: los marcadores lingüísticos de las relaciones sociales, las normas de cortesía, las expresiones de la sabiduría popular, las diferencias de registro, el dialecto y el acento. Finalmente, la competencia pragmática se refiere al conocimiento que posee el candidato acerca de la competencia discursiva: organización, estructuración y ordenación de los mensajes, de la funcional: realización de las funciones

[6] También resulta muy útil la descripción de las diferentes características que debe presentar un examen que hace en el apartado 3 de la parte 1, págs. 34-60, y que resume en el cuadro esquemático de las págs. 49 y 50.

comunicativas en los mensajes, y de la organizativa: secuenciación de los mensajes según esquemas de interacción y de transacción.

3.3. Redacción de los objetivos generales

Para determinar los objetivos generales de una prueba de dominio es conveniente tener en cuenta la sección 5.2.1 del *Marco de referencia*, en la que describen las seis subcompetencias que configuran la competencia lingüística: léxica, gramatical, semántica, fonológica, ortográfica y ortoépica. No obstante, como el *Marco de referencia* no aporta material lingüístico pertinente, se hace necesario consultar el *Plan curricular*. Sin embargo es conveniente tener en cuenta que la relación entre las competencias descritas en el *Marco de referencia* y los inventarios del *Plan curricular* no siempre es unívoca. Hay competencias del *Marco de referencia*, como sucede, por ejemplo, con la competencia semántica, de las que encontramos rastro en diversos inventarios del *Plan curricular*, por ejemplo, en el que presenta en el capítulo 2 de la gramática, en el 6 de las tácticas y estrategias pragmática y en el 9 de las nociones específicas (Navarro y Navarro 2008: 7).

3.4. Redacción de los objetivos específicos

El *Marco de referencia* utiliza una serie de categorías o parámetros para describir lo que los candidatos deben ser capaces de realizar. Para la redacción de los objetivos específicos de las especificaciones de un examen es aconsejable consultar la sección 4.4. del *Marco de referencia*, en la que se detallan las actividades comunicativas de lengua y las estrategias que debe ser capaz de realizar un candidato. El *Marco de referencia* diferencia entre actividades y estrategias de expresión (oral y escrita), de comprensión (auditiva y de lectura), de interacción (oral y escrita) y de mediación (también oral y escrita).

En concreto, para elaborar los objetivos específicos de una prueba de Expresión escrita es conveniente consultar las secciones 4.4.1.2. "Actividades de expresión escrita" y 4.4.3.2 "Actividades de interacción escrita". Los objetivos se pueden enunciar de manera general: "Escribe frases y oraciones sencillas enlazadas con conectores sencillos como "y", "pero" y "aunque"" (Consejo de Europa 2002b: 64) o "Escribe notas breves y sencillas sobre temas relativos a áreas de necesidad inmediata" (Consejo de Europa 2002b: 82) o específica: "Es capaz de escribir una serie de frases y oraciones sencillas sobre su familia, sus condiciones de vida, sus estudios, su trabajo presente o el último que tuvo" (Consejo de Europa 2002b: 65) o "Escribe cartas personales muy sencillas en las que da las gracias o se disculpa" (Consejo de Europa 2002b: 82).

3.4.1. Las tareas de un examen de dominio

Según la definición que proporciona el *Marco de referencia* (155-165) en el capítulo 7, la tarea es una acción intencionada que se realiza en un ámbito, con una intención claramente definida y un resultado específico. Para realizarla es preciso activar de manera estratégica las competencias especificas. En un examen de dominio se recogen muestras de actuación de los candidatos mediante la realización de pruebas que integran tareas estandarizadas, representativas de las que un usuario de la lengua del nivel que se está evaluando realizaría en la "vida real". Para su clasificación se puede seguir la definición que se hace de los objetivos generales en el capítulo 1 de los Niveles de referencia del español.

Para diseñar las tareas de una prueba de dominio es imprescindible tener en cuenta el contexto en el que dicha tarea se va a producir (sección 4.1, *Marco de referencia*: 48). Es esencial conocer el perfil de la población candidata a la que se dirige la prueba con el fin de determinar los ámbitos en que se organiza la

vida social, las situaciones externas que surgen en cada uno de los ámbitos, las condiciones y restricciones, los contextos mentales de los interlocutores y los temas de comunicación.

De los cuatro ámbitos o esferas de acción en que se organiza la vida social: personal, público, profesional y educativo, es preciso determinar cuáles resultan primordiales para la población candidata meta de cada uno de los niveles con el fin de precisar las situaciones, propósitos comunicativos, tareas, temas y textos que configurarán el examen.

Las tareas del examen se desarrollan en situaciones (sección 4.1.2. del *Marco de referencia*, p. 50) que deben estar en correspondencia con las de la «vida real» y que, tal y como se observa en el Cuadro 5 del *Marco de referencia*, se organizan en torno a los siguientes conceptos: lugares, instituciones, personas, objetos, acontecimientos, acciones y textos.

La comunicación se produce en determinadas condiciones externas que imponen ciertas restricciones tanto en el usuario como en sus interlocutores. Se refieren, en concreto, a condiciones físicas (para el habla y para la escritura), condiciones sociales, presiones de tiempo, etc. En la «vida real» las condiciones físicas en las que se realiza la comunicación influyen en la capacidad de los hablantes para poner en acción su competencia comunicativa. Es preciso asegurar, por un lado, que todos los candidatos que se presentan a una prueba la realizan en las mismas condiciones. Además, es conveniente que las tareas de los exámenes incluyan, en la medida de lo posible, las restricciones que se observan en la «vida real».

Tal y como indica el *Marco de referencia* (sección 4.1.4 y 4.1.5. pp: 54-55), el contexto externo, que es extremadamente complejo para percibirlo en su totalidad, es depurado por el usuario de la lengua por medio de diferentes elementos (el aparato perceptivo, los mecanismos de atención, la experiencia a largo plazo, la clasificación práctica de objetos, acontecimientos, la

categorización lingüística) que influyen en la percepción que este tiene del contexto. Los factores del usuario (las intenciones, la línea de pensamiento, las expectativas, la reflexión, las condiciones y restricciones que controlan las elecciones de acción y el estado de ánimo) condicionan la consideración sobre la relevancia que determina el contexto mental para el acto de comunicación, que está condicionado por cómo se percibe el contexto exterior. Pero en este acto de comunicación también es preciso considerar al interlocutor del usuario; el interlocutor puede reaccionar de manera diferente al usuario al no compartir con este, total o parcialmente, las condiciones y restricciones.

Cuando se diseña un examen, con frecuencia se presta toda la atención a la definición de objetivos, a la especificación de los elementos de las competencias, de las actividades comunicativas de la lengua, de las estrategias y de los textos que se van a evaluar, al diseño de las tareas de examen… Sin embargo, es de capital importancia atender a cómo va a percibir el candidato el contexto en que se inscriben los textos (orales o escritos) de las diferentes pruebas. En relación con el contexto mental del usuario, hay algunos elementos del usuario que no podrá activar el candidato durante la realización de una prueba de dominio como, por ejemplo, la experiencia a largo plazo que afecta a la memoria de su interlocutor. Ciertamente, es considerablemente más complicado considerar el contexto mental del interlocutor. En una prueba de dominio el interlocutor, por regla general, es un completo desconocido para el candidato; este hecho puede entorpecer la superposición o congruencia parcial entre los contextos mentales de candidato e interlocutor, que a su vez puede dificultar que el candidato franquee el vacío de comunicación que ocasiona la necesidad de comunicarse.

Los temas son el centro de atención de las tareas de comunicación que se producen en los discursos, las conversaciones o las redacciones y se encuentran integrados en los diferentes ámbitos. El *Marco de referencia* sigue aquí también la estructuración de Threshold Level (1990) en su capítulo 7,

donde clasifica las categorías temáticas en temas, subtemas y "nociones específicas".

Similar organización presentan las Especificaciones de capacidad lingüística («Puede hacer») de ALTE (Consejo de Europa 2002b, Anejo D, pp. 235-248 y ALTE 2002a), donde se detalla lo que puede hacer un candidato en cada uno de los cinco niveles de ALTE y en cada uno de los tres ámbitos o sectores amplios de la vida social en los que, según ALTE, actúan los agentes sociales y constituyen una de las tres principales áreas de interés para la mayoría de los estudiantes de idiomas: social y turismo, laboral y académico (ALTE 2002a: 4). El *Marco de referencia*, sin embargo, distingue cuatro ámbitos: el personal, el público, el profesional y el educativo. Estas escalas, centradas en el usuario, resultan de gran utilidad como listas de comprobación de lo que pueden hacer los candidatos y definir la etapa en la que se encuentran. No obstante, a pesar de la diferencia reseñada, los contenidos son comparables.

3.4.2. Actividades comunicativas de la lengua y estrategias

Las tareas comunicativas descritas arriba se realizan por medio de actividades de lengua de carácter comunicativo (Consejo de Europa 2002b: 14-15, 60-61) como la comprensión (auditiva y de lectura), la expresión (oral y escrita), la interacción (oral y escrita) y la mediación (oral y escrita). Aunque en el § 3.4. ya hemos hecho aconsejado la consulta en el *Marco de referencia* de las actividades comunicativas de la lengua, el objetivo de esta sección es ampliar las explicaciones allí dadas.

Ciertamente no todas las actividades son igual de relevantes en los diferentes niveles. Los procesos de comprensión y de expresión son primarios, previos y necesarios para la interacción. Esta, la interacción, supone un paso más. Interactuar no es solo cuando dos interlocutores participan en un intercambio

oral o escrito en el que se alternan la expresión y la comprensión; interactuar supone, por ejemplo, ser capaz de anticipar el final del mensaje de un interlocutor y de preparar una respuesta. Dado que es una actividad que tiene una gran importancia en el uso y en el aprendizaje de una lengua debería tenerlo, por consiguiente, en la evaluación.

Las actividades de mediación tienen como finalidad facilitar la comunicación entre interlocutores que no son capaces de comunicarse entre sí. Aunque si bien es verdad que estas actividades son relevantes en el funcionamiento lingüístico normal de la sociedad, es posible que se sitúen al margen de los objetivos generales de pruebas de dominio como las de los DELE.

La utilización de las estrategias se realiza con el fin de maximizar la eficacia de la comunicación mediante la adopción de una determinada línea de acción. Se trata de un procedimiento que utiliza el interlocutor para culminar con éxito la tarea de comunicación que está realizando de la forma más completa o más económica posible, dependiendo de sus objetivos. Por consiguiente, su utilización no supone, necesariamente, una carencia en el estudiante. Las estrategias son utilizadas, también, con frecuencia por los hablantes nativos.

El uso de las estrategias conlleva la aplicación de los principios metacognitivos de planificación, ejecución, control y reparación de los diferentes tipos de actividades comunicativas que hemos descrito arriba.

3.4.3. Textos

Son el elemento nuclear del acto de la comunicación por medio de la lengua ya que es absolutamente imprescindible para que este se produzca. Se clasifican en función del género discursivo al que pertenecen: de transmisión oral o de transmisión escrita.

Aunque cualquier texto puede transmitirse por medio de cualquier canal de comunicación, ciertamente existe una estrecha relación entre canal y texto. Los textos transmitidos en el discurso hablado transmiten una información fonética significativa que no es posible transmitir en los textos escritos; por el contrario, los textos escritos presentan unas características paratextuales que dependen del medio espacial y que no se encuentran presentes en los textos orales.

3.4.3.1. Elementos de las competencias necesarios para la realización de las tareas

Para especificar los elementos constituyentes de las competencias necesarios para realizar las tareas y actividades requeridas para enfrentarse a las situaciones comunicativas en las que se ven envueltos los candidatos es precisa la consulta del capítulo 5 del *Marco de referencia* junto con las descripciones que figuran en los *Niveles de referencia para el español*.

C A P Í T U L O 4

Las prueba de expresión escrita y los procedimientos de valoración. Problemática de la evaluación mediante jueces

Los tests de desempeño o de ejecución (performance assessment) son un tipo de test estandarizado (Martínez 2010: 85) de uso cada vez más frecuente en la evaluación psicológica y educativa. Este tipo de evaluación se realiza en diversos ámbitos: gimnasia artística (barras paralelas, barra fija, barras asimétricas, barra de equilibrios, anillas, potro de saltos…), natación sincronizada (tampolín, plataforma…), música (valoración de la ejecución de obras musicales), enseñanza de segundas lenguas (redacciones escritas, producciones orales…), etc. En muchas de las situaciones anteriores expuestas las calificaciones obtenidas por los examinados tienen altas consecuencias. En las disciplinas deportivas una décima arriba o abajo en una ronda de puntuaciones puede suponer ganar o perder un campeonato; en el caso de la música el examinado, por ejemplo, puede jugarse su carrera musical cuando el tribunal decide si entra o no en una orquesta de renombre; en el caso de la evaluación de una lengua extranjera un candidato puede jugarse su futuro profesional cuando redacta un escrito para una prueba de selección para la universidad o para un examen acreditativo del grado de competencia y dominio de un idioma como el DELE.

El empleo de los tests de desempeño en la evaluación educativa comenzó a introducirse en los años ochenta (Lane y Stone 2006), tras las fuertes críticas que se habían realizado a los formatos de selección múltiple. Actualmente, las pruebas de desempeño se pueden encontrar en la mayor parte de los sistemas de evaluación a gran escala junto con pruebas de respuesta construida única.

El objetivo de este tipo de pruebas es emular "el contexto o las condiciones en las que se aplican los conocimientos y destrezas que se intenta evaluar" tal y como se describe en los Standards for Educational and Psychological Measurement en 1999 (Lane y Stone 2006: 388). Aunque los principios para el diseño y elaboración de este tipo de tests son similares a los de otras formas de evaluación (Fitzpatrick y Morrison 1971), por medio de los tests de ejecución es posible evaluar aspectos importantes que no pueden medirse por medio de los de respuesta preseleccionada. Las evaluaciones de desempeño tienen alta fidelidad (Fitzpatrick y Morrison 1971), son transparentes (Frederiksen y Collins 1989) y útiles (Linn, Baker y Dunbar 1991). Y comparadas con otros tipos de tests estandarizados, las evaluaciones de desempeño se caracterizan por su alto grado de autenticidad, mayor complejidad cognitiva, cobertura en profundidad, respuesta elaborada por el propio candidato, mayor coste (Martínez 2010: 87) y proporcionan una medida más directa de los logros de los estudiantes (Lane y Stone 2006: 389). Además, las evaluaciones de desempeño permiten dar oportunidades de elección a los participantes en la prueba (optar entre dos opciones propuestas) en una tarea de escritura (Baker, O'Neil, y Linn 1993; Baron 1991). La prueba de EIE que forma parte del examen del nivel A2 objeto del análisis es un ejemplo de este tipo de test.

El diseño de una prueba de desempeño puede basarse en dos tipos de enfoques: basados en el constructo o en la tarea (Messick 1994). Los diseños fundamentados en el constructo se centran en identificar un completo conjunto de conocimientos, habilidades y otras características que deben ser evaluadas, y determinar qué actuaciones o comportamientos deben ser tenidos en cuenta en la evaluación. Los que se basan en las tareas suelen ser más apropiados para evaluar constructos tales como los musicales o los de la creación artística en los que resulta complicado especificar de antemano los conocimientos y habilidades. En consecuencia, Messick (1994: 22) sugiere que "para diseñar una evaluación de desempeño, cuando sea posible es preferible

utilizar un enfoque basado en un constructo en lugar de usar uno centrado en la tarea" (citado en Lane y Stone 2006, p. 390).

El objetivo ideal que persigue un proceso de calificación es que la calificación obtenida por los candidatos esté en relación con los objetivos que promueve el *Marco de referencia*. Es decir, que cuando se califica como apto a un candidato en uno de los niveles descritos en el *Marco de referencia* es preciso tener la seguridad de que el candidato cumple con los descriptores "puede hacer" del nivel correspondiente que se presentan en el anejo D y que fueron elaborados por ALTE.

4.1. Sistema de desarrollo de las especificaciones del test y del diseño de tareas

El proceso de desarrollo de un examen de dominio es un proceso cíclico e iterativo que comienza con la delimitación del marco conceptual y que incluye una descripción del constructo que debe evaluarse, las consecuencias de la evaluación y las inferencias que pueden realizarse a partir de los resultados de la evaluación. La construcción de esta teoría sirve de guía para el desarrollo del sistema evaluativo. En las especificaciones del examen deben figurar los contenidos de los procesos cognitivos, la características psicométricas de las tareas, así como cualquier otra información que resulte pertinente para el proceso evaluativo (Lane y Stone 2006: 390). Posiblemente la guía más detallada en relación con este tema es la publicada en 2009 por el Consejo de Europa y producida por ALTE bajo la dirección de Michael Milanovic, una nueva versión de otra anterior publicada en octubre de 2002 (ALTE 2002b).

La finalidad de la guía es servir de orientación a las entidades certificadoras y a los responsables de las mismas. En ella se proporciona una detallada descripción de los diversos pasos que deben seguirse para diseñar y elaborar un examen. Primeramente ofrece unas consideraciones generales acerca de

cómo definir el nivel de dominio de un examen y la manera de utilizar el *Marco de referencia*, de la validez, la fiabilidad, la ética y la justicia y la planificación del trabajo.

Seguidamente trata aspectos relacionados con el desarrollo del test: planificación, diseño, experimentaciones y documentos que hay que elaborar para las partes interesadas. También se refiere al sistema de calificación, con información relativa a los criterios y escalas de calificación de las pruebas de Expresión oral y escrita.

A continuación se centra en la construcción del test: producción de materiales y control de calidad; la siguiente fase es la de administración del test, en la que es preciso detallar objetivos y proceso.

Finalmente, es necesario considerar el formato de los documentos en los que se presenten los resultados de los análisis realizados y la monitorización de las escalas de revisión.

Este proceso es recurrente y, en consecuencia, si se detectan problemas en la etapa de seguimiento es posible que haya que modificar las especificaciones y volver a continuar el proceso anteriormente seguido. El objetivo final es que el examen construido sea fiable y que el uso que se haga de sus resultados sea válido, es decir, que mida de manera consistente lo que se quiere medir y que las conclusiones que se extraigan de los resultados de la prueba sean significativas, apropiadas y útiles.

Las tareas que constituyen una prueba de desempeño deben representar de manera sistemática la intención del constructo y además de estar estandarizadas, deben ser válidas, fiables y equitativas. Para ello es preciso contar con unas especificaciones cuidadosamente desarrolladas con el fin de que cada tarea que integre la prueba sirva para medir algo único y diferente de las otras tareas y garantice la comparabilidad de resultados en el tiempo

(Haertel y Linn 1996 y American Educational Research association, American Psychological Association y National Council on Measurement in Education 1999).

Las especificaciones de una prueba deben reflejar "el contenido, la naturaleza y el nivel de los procesos cognitivos que deben evaluarse [… e] incluir información sobre la ponderación de los contenidos y de los procesos cognitivos, así como de las propiedades psicométricas de las tareas (Lane y Stone 2006: 391). Las especificaciones también deben informar acerca de los procedimientos de administración de la prueba: instrucciones de administración, duración, procedimientos de puntuación, criterios de calificación…, y tienen que servir de guía para el diseño de las tareas, junto con el propósito de la evaluación, la población meta y la previsión de las inferencias de resultados.

En el diseño de las tareas de los diversos tipos de evaluación, incluida la de desempeño, es preciso tener en cuenta el propósito que se persigue, la población candidata y la interpretación que se hará de los resultados del test. Desde los años 90 diferentes investigadores han establecido directrices generales para el diseño de las tareas de una evaluación de desempeño. Barón (1991), por ejemplo, argumentó que las tareas deben ser ricas en contenidos, valoradas por expertos en el dominio evaluado, profesores y estudiantes y reflejar los avances actuales en teorías de cognición, aprendizaje, instrucción y motivación. Otros autores, como Marzano, Oickering y McTighe (1993), proporcionan modelos específicos para el desarrollo de evaluaciones de desempeño:

- Identificación de una norma de contenido.
- Estructurar la tarea en torno a un pensamiento complejo.
- Redacción del primer borrador de la tarea basada en la información de los dos primeros pasos.

- Incorporación de la categoría de procesamiento de la información.
- Incorporación de los estándares de las categorías de comunicación.

Una vez concluidas las etapas de desarrollo, en un sistema de evaluación certificativa comienza, por lo general, el sistema de producción de exámenes.

4.2. Vinculación de los exámenes con el *Marco de referencia*

Vincular un sistema de certificación con el *Marco de referencia* es un proceso que requiere de un trabajo minucioso realizado por especialistas en enseñanza y evaluación. Para ello es preciso seguir las especificaciones del *Manual.*

Las fases de un proceso de vinculación de un test con el *Marco de referencia* son las siguientes:

1. Fase de familiarización. El seguimiento de las especificaciones del *Manual* correspondientes a esta fase garantizará que las personas involucradas en la elaboración de los exámenes y su vinculación con el *Marco de referencia* tienen interiorizado el contenido del mismo, en especial, los descriptores ilustrativos asociados a las escalas de los niveles comunes de referencia.
2. Fase de especificación del contenido del examen. Esta fase consiste en el análisis y la descripción, tanto general como detallada, del contenido del examen, en términos de las categorías del *Marco de referencia.* La descripción y el análisis se centran en: la situación o contexto de examen, la cobertura o sílabo, el proceso de elaboración del examen, las tareas que incluye y el sistema completo de calificación. Es necesario justificar el desarrollo de esta fase y dejar constancia documental del proceso mediante los correspondientes informes y la representación gráfica de la relación con los niveles, tal y como se propone en el *Manual.*
3. Fase de Estandarización. El objetivo de esta fase es la interpretación consensuada y compartida del significado de los niveles comunes de

referencia del *Marco de referencia* por parte de todas las personas implicadas en el desarrollo del examen. La estandarización requerirá la participación de profesionales expertos, familiarizados con los niveles del *Marco de referencia*, con tareas e ítems previamente calibrados y con muestras ilustrativas de los niveles. En relación con las tareas e ítems de las pruebas de comprensión, el grupo de expertos debe calibrar su nivel de dificultad en relación con los niveles del *Marco de referencia*. Por lo que respecta a las tareas de expresión e interacción, es preciso trabajar con muestras orales y escritas de candidatos. Esta fase del proceso culminará con la realización del informe correspondiente, que debe elaborarse siguiendo las pautas del *Manual*.

4. Fase de Validación a través del análisis de los datos del examen. Validación psicométrica de los resultados obtenidos con el procedimiento de estandarización. El resultado de la fase de validación deberá documentarse, al igual que las anteriores.

La documentación resultante del proceso completo de vinculación de los exámenes DELE con el *Marco de referencia* deberá ajustarse a las especificaciones del Capítulo 7 del *Manual*, correspondientes a la modalidad de "Informe detallado, dirigido a especialistas".

4.3. Diseño y elaboración de los criterios de calificación

Uno de los aspectos críticos de los tests de desempeño es la correcta asignación de las puntuaciones a las tareas realizadas. Para ello es preciso disponer de criterios de calificación (scoring rubrics) en los que se detallen tanto los criterios de valoración de las respuestas como los procedimientos de puntuación (Clauser 2000). La elaboración de este documento es uno de los hitos en el proceso de diseño de una prueba de desempeño. El objetivo que se persigue con esta guía es que las puntuaciones que los examinadores otorguen a los candidatos tras consultarla sean consistentes e invariantes a través de calificadores, tareas, convocatorias, etc. Guilford (1954: 263-278)

estudia los tipos de escalas de calificación existentes y sus ventajas y desventajas. Lógicamente, para alcanzar esta meta y conseguir niveles adecuados de fiabilidad es aconsejable que los examinadores asistan regularmente a cursos de formación.

En el Anejo A del *Marco de referencia* (pp. 197-207) se detallan las siguientes las orientaciones para el desarrollo de los descriptores de las escalas de calificación:

- La formulación de los descriptores sea positiva.
- La redacción debe de ser precisa para evitar problemas de diversas interpretaciones de los mismos.
- Se debe evitar la jerga técnica en la descripción de los descriptores con el fin de que estos sean transparentes, breves, independientes y claros.

Para desarrollar unas escalas de calificación en las que las descripciones de un grado concreto se relacionen con niveles concretos de dominio lingüístico, en el Apéndice A del *Marco de referencia* (199) se indica que es posible utilizar tres tipos de métodos: intuitivos, cualitativos y cuantitativos, aunque se añade que los mejores procedimientos de elaboración de escalas "combinan los tres enfoques en un proceso complementario y acumulativo".

4.3.1. Fase intuitiva

El proceso de redacción de las escalas de calificación es un proceso básicamente intuitivo en el que es preciso decidir el sistema de redacción del borrador de escalas de calificación que se va a seleccionar. La definición de cada una de las bandas varía dependiendo de si lo que se evalúa es un producto o un proceso, de las exigencias de las tareas, de la población meta, del propósito de la evaluación y de la interpretación que vaya a hacerse de los resultados previstos. Las posibilidades son:

- Pedir a un experto que escriba la escala consultando las fuentes que tenga a su alcance.
- Seleccionar a varios especialistas (comisión) para que desarrollen los borradores y los comenten sobre la base de su experiencia o comparando con alumnos o con muestras de actuación. Este método ha sido criticado por algunos autores (Gipps 1994 y Scarino 1996, 1997).
- Ampliar la duración del método anterior estableciendo un proceso en el que los especialistas, tras alcanzar un conocimiento compartido de criterios y niveles, comprueben y utilicen la información obtenida para modificar la redacción hasta que se alcanza un quórum.

Para sistemas de evaluación a gran escala, Lane y Stone (2006) indican que en el desarrollo de los criterios de calificación es aconsejable contar con un grupo de especialistas en el contenido de las pruebas y con amplia experiencia como educadores. En este sentido, es necesario considerar cómo habría que realizar la selección de los integrantes del equipo que se encargará de la elaboración del documento, cuántos miembros debería tener, el número de pruebas que tendrá que calificar cada examinador, la formación que deberían tener y qué habilidades, conocimientos sobre el tema y experiencia tendrán que acreditar.

El equipo, en un primer momento, tendrá que realizar actividades de familiarización con el *Marco de referencia* con el fin de que todos los participantes en el proceso conozcan sus escalas. Los integrantes del equipo tendrán que determinar si prefieren diseñar una guía holística, en la que los examinadores únicamente tengan que emitir un único juicio global acerca de la calidad del texto escrito por el candidato; una analítica, en la que las descripciones del desempeño se dividan en partes (aspectos cualitativos); o una en la que se valore el rasgo primario, es decir, lo bien o mal que los candidatos han escrito en un espectro del discurso que se ha acotado entre ciertos límites (persuasión, explicación…) (Huot 1990, Miller y Crocker 1990, Mullis 1984 y Weige 2002).

Las guías holísticas suelen recomendarse más para tareas relativamente simples (Johnson, Penn y Gordon 2009; Arter y McTighe 2001); las analíticas, sin embargo, parecen ser más adecuadas para tareas de desempeño más complejas con múltiples elementos (Welch 2006), mientras que la calificación del rasgo primario se utiliza para focalizar la atención en una pauta especificada con anterioridad. En los capítulos 4 y 5 del *Marco de referencia* se pueden encontrar las escalas ilustrativas de una serie de categorías que se relacionan con el esquema descriptivo que se presenta en estos capítulos. Sin embargo, no es posible utilizar todas las escalas en todos los niveles y es preciso reducir su número. El propio *Marco de referencia* (193) indica que, según la experiencia, "más de cuatro o cinco categorías comienzan a provocar una sobrecarga cognitiva, y que siete categorías es psicológicamente un límite máximo". La validez en la interpretación y el uso de las calificaciones dependen de la interrelación entre el constructo que se está midiendo y las puntuaciones obtenidas (Messick 1989).

4.3.2. Fase cualitativa

En caso de que se utilicen métodos cualitativos y cuantitativos en el mismo proceso es posible comenzar el proceso de dos modos diferentes, con descriptores o con muestras de actuación (Consejo de Europa 2002b*:* 199).

Si se parte de los descriptores, es preciso analizar aquello que se va a describir y redactar después borradores de descriptores de los atributos concretos. Si se parte de muestras de actuación es necesario comenzar con muestras representativas de las actuaciones para que los examinadores indiquen lo que ven cuando trabajan con las muestras.

En caso de que se haya elegido la primera de las posibilidades (partir de descriptores) es posible realizar pequeños talleres de trabajo en los que se fragmente el borrador de escalas y se pida a los grupos de informadores que

reordenen las escala, justifiquen su decisión e identifiquen los elementos clave que les ayudaron a hacerlo o les confundieron (procedimiento empleado para desarrollar las escalas de *Eurocentres*). Otra posibilidad es pedirles a los informadores que organicen los descriptores en conjuntos teniendo en cuenta los atributos que describen y los niveles (Pollit y Murray 1996).

La elección de la segunda de las opciones (muestras de actuación) hace necesario utilizar alguna de las siguientes variantes:

- Comparación de descriptores con actuaciones típicas de los niveles de las bandas de calificación (Alderson 1991).
- Ordenar las actuaciones según su nivel y describir qué aspectos son los que justifican la clasificación realizada con la intención de definir el rasgo que determina la ordenación (Mullis 1981).
- Una variante de la anterior es debatir acerca de los límites entre niveles con las muestras previamente ordenadas, con el fin de determinar sus características clave y formular con cada una de ellas una pregunta breve de criterio con una respuesta de sí o no (Upshur y Turner 1995).

A continuación es preciso detallar los rasgos o aspectos detallados de las respuestas. Por lo general, los distintos rasgos se puntúan por medio de escalas tipo Likert con varios grados (Likert 1932, Martínez 2010: 88). El número de puntos con que se valora cada atributo deberá ser lo suficientemente extenso como para permitir diferenciar entre niveles de rendimiento, pero no será tan amplio que dificulte la diferenciación entre ellos (Lane y Stone 2006).

4.3.3. Fase cuantitativa

El objetivo de esta fase es cuantificar el material comprobado anteriormente de manera cualitativa. Para su realización es posible utilizar alguna de las siguientes técnicas:

- Analizar de manera detallada el discurso de un conjunto de actuaciones que ya han sido valoradas previamente, con el fin de determinar y contabilizar la incidencia de distintas características cualitativas. Para determinar cuáles son las características relevantes para determinar la clasificación realizada por los examinadores se utiliza la regresión múltiple (Fulcher 1996).
- Valorar las actuaciones mediante una escala analítica de varios atributos y determinar qué categorías fueron más decisivos para determinar el nivel (Chaloub-Deville 1995).
- De entre la familia de medidas o de modelos de elaboración de escalas de medición que ofrece la TRI el más potente es el modelo de Rasch, propuesto en 1960 por el matemático danés Georg Rasch. Las ventajas del modelo de Rasch respecto a la TCT y a otros modelos TRI han sido ampliamente difundidas (Prieto y Delgado 2003):
  - Medición conjunta: los parámetros de los candidatos y de los ítems se expresan en las mismas unidades y se localizan en el mismo continuo, lo que permite analizar las interacciones entre los candidatos y los ítems.
  - Objetividad específica: una medida solo puede ser considerada válida y generalizable si no depende de las condiciones específicas con que ha sido obtenida. Es decir, las medidas de los examinados son estimaciones independientes de la severidad de los calificadores y de la dificultad de las tareas.
  - Propiedades de intervalo: la interpretación de las diferencias en la escala es la misma a lo largo del atributo medido, es decir, a diferencias iguales entre sujeto e ítem le corresponden probabilidades idénticas de una respuesta correcta.
  - Especificidad del error típico de medida: la objetividad específica no implica que la precisión de las estimaciones de los parámetros sea similar en distintos conjuntos de ítems y de personas. Por ejemplo, si los candidatos son de alto nivel, se estimarán con más precisión los parámetros de los ítems difíciles.

Una de las aplicaciones relevantes del modelo de Rasch es determinar la dificultad de los ítems en relación con el nivel de competencia de los candidatos, lo que permite graduar los ítems en la misma escala. Un desarrollo

del enfoque permite utilizarlo para escalonar descriptores de dominio comunicativo de la lengua.

La ventaja de un análisis de Rasch es que proporciona una medición independiente de la muestra (de los examinados o de los ítems). Es decir, las puntuaciones de los candidatos son independientes de los ítems administrados, y los valores de los ítems son independientes de la muestra de candidatos empleada en la calibración. El análisis de Rasch se puede emplear de diversas formas para valorar descriptores por escalas: (i) asignar valores numéricos a los datos obtenidos en la fase cualitativa; (ii) es posible elaborar pruebas para hacer operativos los descriptores de dominio de la lengua en ítems concretos de pruebas y posteriormente esos ítems se pueden escalonar con el análisis de Rasch y tomar sus valores en la escala para indicar la relativa dificultad de los descriptores (cf. Kirsch y Mosenthal, 1995); y (iii) pueden utilizarse los descriptores como ítems de pruebas para la evaluación que el profesor realiza a sus alumnos y de este modo graduar los descriptores directamente en una escala aritmética.

Además de su utilidad para desarrollar escalas, el método de Rasch también puede utilizarse para analizar las formas en que se utilizan las bandas de calificación (cf. Milanovic y Saville 1996 y Tyndall y Kenyon 1996).

4.4. Estudio de las fuentes de error

Las puntuaciones que obtienen los candidatos que han realizado una prueba de desempeño no dependen únicamente del nivel de los examinados en el constructo de interés. La utilización de las escalas de calificación se basa en la suposición de que un examinador es capaz de realizar una correcta observación cuantitativa con precisión y con un cierto grado de objetividad (Guilford 1954: 278). Guilford aconsejaba analizar las diversas fuentes de

error que pueden poner en peligro la calidad de un proceso de calificación. Popham (1990) siguió la recomendación de Guilford y proporcionó un marco conceptual útil para la organización de estos factores. Las tres posibles fuentes de error que es necesario monitorear cuidadosamente son las siguientes: la utilización de las escalas de calificación, el proceso de calificación y los examinadores.

4.4.1. Utilización de las escalas de calificación

Un problema que pueden tener los calificadores cuando consultan las escalas de calificación es que no tengan claro qué es lo que tienen que calificar si el atributo no está definido con claridad. También es necesario que la redacción de las escalas sea clara y que se diferencien de manera nítida las descripciones de cada una de las bandas de calificación entre sí. Una terminología confusa o demasiado ambigua puede afectar negativamente al proceso de calificación.

4.4.2. Procedimiento de calificación

También pueden ser fuente de errores intrajuez factores como el cansancio o la fatiga, la prisa, el estado de ánimo, las preocupaciones teóricas en relación con el constructo evaluado o la presión excesiva por las altas consecuencias que puede tener la evaluación realizada. Es posible, asimismo, que a lo largo del proceso de calificación varíe la sensibilidad del calificador ante determinados errores si se detectan repetidamente en los textos de los candidatos. Al principio, un determinado error puede resultar muy relevante para el examinador, pero después de encontrarlo repetidamente puede parecerlo menos. Realmente estas circunstancias son difíciles de detectar. Myford y Wolfe (2004a: 466) llegan a afirmar que es posible que algunos examinadores puedan calificar mejor a determinadas horas del día.

Respecto a la influencia que pueden tener en el proceso de calificación los factores del entorno es posible destacar el tiempo de que disponga el examinador para calificar los exámenes que tiene asignados, la hora del día, las condiciones del lugar en el que realiza el proceso de calificación, el sistema de visualización de las tareas (pruebas originales, fotocopias, escaneadas y visualizadas en la pantalla de un ordenador...). En relación con este último aspecto, están por analizar las consecuencias que pueden tener en la calidad de las calificaciones de los examinadores cuestiones como la visualización de las tareas en la pantalla de un ordenador, la imposibilidad de hacer anotaciones en las tareas de los candidatos al leer los exámenes en una pantalla, etc. Autores como Myford y Wolfe (2004a) han indicado que hay examinadores que pueden experimentar fatiga visual y dolores de cabeza, otros pueden tener dificultades para adaptarse a las exigencias de los sistemas de calificación en línea o pueden experimentar dificultad para concentrarse ante la pantalla de un ordenador.

4.4.3. El examinador como fuente de error

Es preciso estar alerta ante la presencia de sesgos en los juicios de los calificadores. Para analizar su variabilidad es preciso tener en cuenta facetas como la dificultad de las tareas, la severidad o benignidad del calificador y el uso que este haga de los atributos. De hecho, si la dificultad de las tareas es la adecuada y el sistema de calificación ha sido correctamente definido, el comportamiento de los calificadores puede constituir la principal fuente de varianza de las calificaciones. El acto de calificar a los candidatos, señala Cronbach (1990: 584) es un "proceso cognitivo complejo y propenso a errores". El comportamiento de los calificadores es difícil de controlar tanto por los diseñadores de pruebas como por los de escalas de calificación y supone una fuente importante de varianza en las calificaciones (Lane y Stone 2006). El proceso de calificación Las diferencias entre los calificadores en la

asignación de las bandas de calificación, en su severidad o benevolencia/indulgencia, la tendencia central, el efecto de halo o el sesgo a la hora de aplicar las puntuaciones a grupos de diferente género o nacionalidad, etc., "contribuyen al error de medida, a la validez y a la justicia de las evaluaciones." (Prieto 2011: 233). Y no se trata necesariamente de un problema en la formación de los calificadores aunque, ciertamente, sin una buena preparación de los jueces es más fácil que dichos problemas se incrementen. Hay estudios que demuestran que jueces competentes que califican a los mismos sujetos pueden diferir en sus puntuaciones (cf. Watts y Watts y García 2006; Cuxart, Martí y Ferrer 1997; Grossman y Wood 1993) al hacer hincapié en contenidos diferentes al evaluar las tareas.

4.4.4. Procedimientos para prevenir los errores de los calificadores

En primer lugar es imprescindible que en un equipo de calificadores todos sus integrantes estén sólidamente formados y la mayoría deben tener experiencia en la calificación que van a realizar. En determinados casos los examinadores deben acreditar que están libres de prejuicios o creencia positivas o negativas a la hora de enfrentarse a la calificación de las tareas. Pensar que los candidatos de determinadas nacionalidades sistemáticamente tienen que tener buenas o malas calificaciones, que siempre cometen determinados tipos de errores puede influir negativamente en la calificación realizada. Existen programas de formación de calificadores en los que se les proporciona información acerca de los errores más frecuentes entre los examinadores y la manera de evitarlos (Cursos Internacionales de la Universidad de Salamanca e Instituto Cervantes, 2015).

Puede resultar interesante aplicar una serie de estrategias para controlar el comportamiento de los calificadores en tiempo real, cuando estos están calificando pruebas. Los sistemas de calificación en línea permiten que un supervisor de calificación o calificador experto califique aleatoriamente tareas

calificadas por los calificadores con el fin de verificar que los examinadores utilizan las escalas de calificación de manera adecuada. También pueden recibir periódicamente información estadística del comportamiento de los caliricadores con el fin de identificar errores intraevaluador y tratar de efectuar cambios en el comportamiento de los mismos.

Los principales errores de calificación que son atribuibles a factores intrajuez son (seguimos a Myford y Wolfy (2004a: 471) en la clasificación: severidad/benignidad, tendencia central, efecto de halo y restricción de s alcance.. Sin embargo, también hay otros, aunque son menos frecuentemente mencionaos: falta de precisión, error lógico, error de contraste, influencia en los calificadores de sesgos, creencias, actitudes y características de la personalidad, influencias de las características del calificador, error de proximidad, novedad (o primacía) del error y efectos del orden.

4.4.4.1. Severidad/benignidad

El término "benignidad" (*leniency*) parece que fue utilizado por vez primera por Kneeland (1929: 356) para describir la tendencia de un examinador a calificar muy por encima del punto medio de las escalas utilizadas. A Ford (1931) se le atribuye el primer uso del término "severidad" (*severe*) para describir el otro extremo del *contínuum*, es decir, cuando un evaluador que tiende a calificar por debajo del punto medio de la escala.

Lane y Stone (2006) señalan que la benignidad se produce cuando un calificador otorga puntuaciones demasiado elevadas en relación con el nivel de competencia que tienen un candidato, mientras que la severidad se caracteriza por una excesiva exigencia por parte del calificador, lo que ocasiona que las calificaciones dadas a un candidato sean inferiores a lo que le

correspondería. Cronbach (1990) asegura que es el error más grave que un calificador puede introducir en un procedimiento de calificación ya que las puntuaciones de los examinadores siempre deberían estar relacionadas con el nivel de competencia de los candidatos.

Guilford (1954) afirma que un examinador será benévolo independientemente del atributo que esté midiendo. Investigadores, como Schriesheim, Kinichki, y Schriesheim (1979) han profundizado en el estudio de este error y lo han considerado como una característica estable del calificador, como un elemento propio de su personalidad (citado en Myford y Wolfe 2004a:472). Se trata, según Thorndike y Hagen (1977), de una tendencia humana de evitar hacer juicios desfavorables de nuestros semejantes, especialmente cuando el calificador se identifica con el sujeto que está siendo calificado (cf. Myford y Wolfe 2004a: 472).

Autores como Saal, Downey y Lahey (1980) han identificado diversos enfoques para detectar la benignidad/severidad en un grupo de examinadores: (i) comparar las puntuaciones medias de los atributos con los promedios de las escalas de calificación empleadas por los calificadores; (ii) utilizar un análisis de la varianza (ANOVA) para contrastar la significación de las diferencias entre el promedio de las calificaciones de un calificador y los promedios del resto de los calificadores; y (iii) examinar el grado de asimetría de la distribución de frecuencias de las calificaciones en los atributos.

Resulta complicado atribuir la severidad o benevolencia de los calificadores a factores concretos, aunque Eckes (2011: 55) destaca una serie de autores que han profundizado en este sentido: Eckes (2008); McManus, Thompson y Mollon (2006); Myford Marr, y Linacre (1996); Stone (2006) y Landy y Farr (1980). Las causas que pueden influir en que un examinador valore con mayor o menor severidad la actuación de un candidato pueden ser muy variadas: la cantidad de tareas que tenga que calificar y el tiempo de que disponga para hacerlo, factores idiosincrásicos como la personalidad del calificador y su

actitud ante el proceso de calificación en el que va a participar, la experiencia que tenga … En relación con este último aspecto, comenta Eckes (2011: 55) "in some circumstances the most experienced or senior rater may also be the most severe. That rater may feel that he or she must «set the standard» for tha other raters by noticing even small flaws in examinee performance thare otherwise likely to be overlooked". Indica también Eckes que la formación de los calificadores generalmente no resulta eficaz para reducir las diferencias de severidad entre calificadores. McNamara (1996: 127; citado en Eckes 2011:55) recomienda a este respecto "to accept variability in stable rater characteristics as a fact of life which must be compensated for in some way", aunque, según nuestra experiencia (según los datos analizados por Prieto 2015a), habitualmente son los nuevos calificadores los que se muestran excesivamente severos o benévolos, mientras que los calificadores veteranos, que tienen más experiencia y han realizado cursos de formación, presentan unos valores de severidad no tan extremos.

Para poder reducir las diferencias existentes entre los calificadores McNamara recomienda utilizar la información que facilita el modelo MFRM en relación con las tendencias de severidad o benevolencia que presentan los calificadores, para analizar en profundidad la compleja conducta del calificador.

Para minimizar el efecto de la benevolencia / severidad en un equipo de calificadores se han propuesto diversas estrategias (Myford y Wolfe 2004a: 473):

- Ayudar a los evaluadores, especificando claramente las definiciones de los atributos y, si es posible, proporcionar descripciones y ejemplos de anclaje para cada una de las categorías de la escala, de forma que el examinador tenga una idea clara de lo que significa cada una de ellas y sea capaz de distinguir entre las diversas bandas de calificación de los atributos.

- Pedir a los calificadores que elaboren escalas de calificación que tengan bandas de calificación tanto en el lado positivo como en el negativo con el fin de contrarrestar las tendencias de determinados examinadores a ser excesivamente benévolos en sus calificaciones. De este modo, los calificadores podrán diferenciar con facilidad entre los distintos niveles de rendimiento a lo largo del *contínuum*.
- Entrenar a los calificadores acerca de las consecuencias que una excesiva benevolencia o severidad puede tener en sus calificaciones con el fin de que traten de evitar esta tendencia.
- Pedir a los calificadores que ordenen a los candidatos en función de su competencia con el fin de que se vean obligados a discriminar entre ellos.
- Realizar una doble o triple calificación para intentar compensar los efectos que una excesiva benevolencia o severidad puede tener en los candidatos.
- Utilizar técnicas estadísticas para identificar a los calificadores excexivamente benévolos o severos.

4.4.4.2. Tendencia central

La tendencia central consiste en un tipo especial de restricción de rango (Myford y Wolfe 2004a). Se produce cuando un examinador utiliza con excesiva frecuencia la banda de calificación intermedia de cada uno de los atributos y asigna en escasas ocasiones las puntuaciones altas y bajas de la escala. Este tipo de actuación por parte de un calificador ocasiona que en los resultados de sus valoraciones se sobreestime el nivel de competencia de candidatos con bajo nivel de competencia y se minusvalore el de los candiatos con alto nivel.

Son numerosas las definiciones sobre la tendencia central. Landy y Farr (1983) la han definido como "the avoidance of extreme (favorable or unfavorable) ratings or preponderance of ratings at or near the scale midpoint" (citado en Myford y Wolfe 2004: 476). DeCotiis, amplió la

definición utilizando un tono algo más literario "a rater's unwillingness to go out on the proverbial limb in either the favorable or unfavorable direction (citado en Myford y Wolfe 2004: 476) Otros autores como Linn y Gronlund (2000) han indicado que la tendencia a evitar categorías extremas en una escala es un estilo de actuación para algunos calificadores, lo que puede dificultar la corrección de esta tendencia.

(Myford y Wolfe 2004a, 2004b) definen la tendencia central como un tipo especial de restricción de rango que se produce cuando un examinador utiliza con excesiva frecuencia la banda de calificación intermedia de cada uno de los atributos y asigna en escasas ocasiones las puntuaciones altas y bajas de la escala. Este tipo de actuación por parte de un calificador ocasiona que en los resultados de sus valoraciones se sobreestime el nivel de competencia de candidatos con bajo nivel de competencia y se minusvalore el de los candidatos con alto nivel.

4.4.4.3. Efecto de halo

Con el fin de profundizar en la explicación de la variabilidad que se produce en la actuación del calificador, independientemente de la severidad/benignidad, es posible analizar otras tendencias que pueden observarse en la calificación de los examinadores (cf. Barrett 2005, Knoch et al 2007: Myford y Wolfe 2004a, 2004b, Wolfe 2004 y 2009). Dos de las más relevantes son el efecto de halo y la tendencia central.

La expresión *efecto de halo* fue acuñada por Thorndike (1920: 25), aunque se considera a Wells (1907) como el primer investigador que identificó este efecto. En las pruebas de idiomas no se comenzó a analizar este sesgo hasta el último cuarto del siglo pasado por Yorozuya y Oller, Jr. (1980), (cf. Farrokhi y Esfandiari 2011: 1532). El efecto de halo se produce cuando un calificador elige con excesiva frecuencia la misma categoría en los diferentes atributos sin

tener en cuenta el diferente nivel de competencia que puede haber demostrado tener el candidato en cada uno de ellos ni las diferencias existentes. Según Myford y Wolfe (2004a: 474, se trata del error de los calificadores más estudiado y el que ha recibido mayor atención por la literatura científica.

Los atributos que se utilizan para medir la capacidad de los candidatos —nivel en la variable latente— en un sistema de calificación analítico, normalmente están diseñados para representar distintas características de la actuación lingüística que es de interés. Por este motivo, no es sorprendente constatar que las calificaciones de desempeño de los diversos atributos están correlacionadas entre sí, ya que los atributos, aunque conceptualmente diferentes, deben trabajar operacionalmente de manera coordinada para definir el constructo subyacente. Este hecho plantea problemas para la detección del efecto de halo verdadero.

Durante años, los investigadores han propuesto un gran número de definiciones conceptuales de halo (Fisicaro y Lance 1990, Fisicaro y Vance 1994, Saal, Downey y Lahey 1980, King, Hunter, y Schmidt 1980 y Robbins 1989). Fisicaro y Lanza (1990) detallan que el efecto de halo puede deberse a un mínimo de tres procesos cognitivos diferentes:

1. La impresión general de la actuación de un candidato tiene una influencia directa en la puntuación que asigna un calificador en cada uno de los atributos de calificación;
2. La elección de una determinada puntuación en un atributo particularmente destacado (esto puede suceder, por ejemplo, en Corrección, que en este examen se valora junto con el alcance) condiciona la que se asigne al resto de atributos;
3. Un calificador no consigue discriminar adecuadamente entre características de actuación conceptualmente distintas que representan cada una de los atributos analizados.

Según Eckes (2011:66) el efecto de halo se refiere a un sesgo cognitivo por el cual un calificador tiene tendencia a asignar la misma banda de calificación en todos los atributos sin tener en cuenta las posibles diferencias de rendimiento que pueda tener el candidato en cada una de ellos y sin considerar que estos, los atributos, son conceptualmente distintos (Cooper 1981; Saal, Downey y Lahey 1980). Cuando la mayoría de los calificadores presenta efecto de halo, las puntuaciones utilizadas para valorar la actuación del candidato son similares en todos los atributos con que se analizan a los candidatos. La aparente falta de diferencia entre los atributos puede ser un reflejo de la incapacidad de los calificadores para diferenciarlos. Sin embargo, es importante aclarar que la inexistencia de diferencias entre atributos no implica, necesariamente, que los examinadores exhiban efecto de halo.

Se han realizado diversas propuestas con el fin de minimizar los efectos de este efecto. No obstante, dado que se trata de un efecto difícil de erradicar (Cascio 1982: 318), algunos enfoques han tenido un éxito limitado. A continuación exponemos algunas de las estrategias recomendadas para reducir sus efectos:

- Asegurarse de que cada banda de calificación está cuidadosamente definida y que las distinciones entre los atributos son claramente identificables.
- Si la calificación se realiza con atributos diferentes, invertir, en ocasiones, el orden de las bandas de calificación altas y bajas con el fin de evitar que los examinadores asignen sistemáticamente la misma puntuación en atributos conceptualmente diferentes.
- Informar a los examinadores de las consecuencias que el efecto de halo puede tener en sus calificaciones con el fin de que traten de evitar esta tendencia.

- Pedir a los calificadores que evaluén primeramente a todos los candidatos en un único atributo antes de pasar a los siguientes.
- En caso de que haya un gran número de atributos considerar la posibilidad de reducirlos con el fin de no provocar sobrecarga cognitiva entre los examinadores. Si no es posible reducir los atributos, dividir la calificación de estos entre los calificadores, de modo que cada calificador no tenga que trabajar con el total de atributos.

4.5. Formación de los calificadores

Otro aspecto esencial del proceso es la formación de los calificadores. Mediante ella es posible minimizar la existencia de algunos de los sesgos de calificación que se acaban de señalar. En un proceso de formación, según detalla el *Manual*, es preceptivo:

- Llevar a cabo actividades de familiarización con los descriptores de las escalas de referencia del *Marco de referencia* correspondientes al nivel de referencia, en este caso, como se verá más adelante, el A2.
- Trabajar con ejemplos de rendimiento de muestras obtenidas en seminarios de estandarización (Breton, Lepage y North 2008) para conseguir una comprensión adecuada de los niveles del *Marco de referencia*;
- Familiarización con las escalas locales.
- Sesiones de ilustración y práctica controlada con muestras locales y unificación de criterios a partir de informes de calificación con el fin de desarrollar habilidades para relacionar las tareas con los niveles de rendimiento;
- Fase de calificación libre y unificación de criterios para garantizar que todo el equipo de calificadores comparte la misma interpretación y que esta se aplica de forma consistente.

- Evaluación del examinador: calificación individual de muestras de exámenes.

La relación de materiales necesarios para la preparación y desarrollo de las distintas etapas del proceso es prolija. Se detalla a continuación la relación que recomienda el *Manual* (p. 55) en una tabla resumen (tabla 4) para la formación de las destrezas de expresión.

Tabla 4

Formación para la estandarización y la evaluación comparativa: resumen

Actividad	Material necesario	Duración	Integrantes	Sugerencias
FAMILIARIZACIÓN	• Listas de preguntas basadas en recordatorios del MCER (casillas) • Fotocopias de las listas de preguntas • Fotocopias de las tablas 1 y 2 del MCER • Versiones reducidas de la tabla 2 del MCER, otras escalas	2 horas	Coordinador Posibilidad de grupos numerosos	Utilizar el paquete on-line de auto-formación si es posible.
FORMACIÓN (con las destrezas de expresión)	• Vídeos de actuaciones normalizadas (mínimo 8) • Textos normalizados (mínimo 8) Fotocopias de escalas específicas de una destreza: • Tabla 3 del MCER/tablas B1-B3 (actuaciones orales) • Tabla B4 (actuaciones escritas) Fotocopias de: • Hojas de valoración para participantes (formularios B2-B3) • Formularios de valoración para el coordinador (form. B4) Fotocopias de otras escalas complementarias pertinentes	3–4 horas por destreza: 30min: introducción 90min: muestras ilustrativas 90min: muestras locales	Coordinador Máximo 30 personas	Trabajar dos destrezas por jornada o dedicar media jornada a la formación y la otra media a la valoración comparativa con relación a una destreza determinada.
EVALUACIÓN COMPARATIVA DE MUESTRAS DE ACTUACIONES (de expresión)	• Vídeos locales (mínimo 8) • Textos locales (mínimo 8) • Fotocopias de escalas específicas de una destreza: • Tabla 3 del MCER/tablas B1-B3 (actuaciones orales) • Tabla B4 (actuaciones escritas) Fotocopias de: • Hojas de valoración para participantes (formularios B2-B3) • Form. de valoración para el coordinador (form B4) • Fotocopias de otras escalas complementarias pertinentes	3–4 horas por destreza: 30min: introducción 90min: calibración 90min: localización	Coordinador Máximo 30 personas	Trabajar dos destrezas por jornada o dedicar media jornada a la formación y la otra media a la evaluación comparativa con relación a una destreza determinada.

MCER: *Marco de referencia*

Tabla adaptada del *Manual*, p. 55.

En la fase de trabajo con ejemplos de rendimiento, también denominada fase de ilustración, se justifica la calificación con ejemplos en grupos e individualmente y se justifica por qué un candidato no alcanza la banda

superior, con ejemplos. El objetivo que se persigue es que el examinador aprenda a relacionar las tareas con los niveles de rendimiento.

En la fase de desarrollo de habilidades se califican diferentes muestras y se debate entre los asistentes con el fin de comprobar que el calificador es capaz de relacionar las tareas con los niveles de rendimiento.

El objetivo que se persigue en el curso es que al finalizar el mismo el examinador haya internalizado los criterios y las bandas de calificación y desarrollado las habilidades necesarias para ser capaz de clasificar a los candidatos en la escala de niveles del *Marco de referencia*. Por lo general el curso puede centrarse en uno de los "niveles amplios" que especifica el *Marco de referencia* (25): usuario básico, usuario independiente y usuario competente o en uno de los niveles inferiores: nivel *Acceso*, nivel *Plataforma*, nivel *Umbral*, nivel *Avanzado*, nivel *Dominio operativo eficaz*, nivel *Maestría*.

C A P Í T U L O 5

Modelos de análisis de las evaluaciones mediadas por calificadores

Por medio de los tests es posible realizar inferencias y tomar decisiones sobre aspectos importantes de las personas. Lógicamente, es necesario asegurarse que las inferencias son pertinentes con el fin de que no se perjudique a los sujetos a los que se les aplican. Para garantizar que las las decisiones tomadas a partir de ellos son las adecuadas es necesario estimar su fiabilidad y su validez, y para ello se hace preciso utilizar teorías de los tests. Las dos grandes teorías que pueden utilizarse para construir y analizar los tests son la Teoría Clásica de los Tests (TCT) y el enfoque de la Teoría de respuesta a los Ítems (TRI) (Muñiz 2010: 59).

5.1. La Teoría Clásica de los Tests (TCT)

Se trata del enfoque predominante en la construcción y análisis de los ítems (Muñiz 2010: 59-60). Sus orígenes se remontan a los pioneros trabajos de Spearman (1904, 1907, 1913) y hacia mediados del siglo anterior ya se encuentra plenamente desarrollada. Gulliksen fue quien revisó y sistematizó en 1950 todo el corpus teórico relacionado con esa teoría.

La propuesta de Spearman consiste en asumir que la puntuación empírica que un sujeto obtiene en un test (X) está formada por el valor real del atributo la puntuación verdadera de esa persona en el test (V) y un error (e): $S = V + e$, que puede deberse a diversas razones: comportamiento del propio sujeto, el contexto en el que se realizó el test, el propio test… De tal forma que es preciso diferenciar entre el valor real del atributo medido (puntuación

verdadera) y la medida falible que se obtiene en el proceso de medición (puntuación observada).

Spearman desarrolló un modelo formal denominado *Modelo Clásico* o *Modelo Lineal Clásico*, fundamentado en varios supuestos a partir de los que se definen los conceptos de puntuación verdadera y de error y se extraen consecuencias acerca de su aplicabilidad práctica para cuantificar los errores y corregir su efecto (Abad, Olea, Ponsoda y García 2011: 76).

- Primer supuesto: definir la puntuación verdadera (V) como la esperanza matemática de la puntuación empírica: $V = E(X)$. De modo que se define la puntuación verdadera de un sujeto en un test como la puntuación que se obtendría como media se se aplicase infitinas veces el test.
- Segundo supuesto: se asume que no existe relación entre los valores de las puntuaciones verdaderas de los sujetos, sean estos altos o bajos, y el tamaño de los errores que afectan a estas puntuaciones: $r\,(v,e) = 0$. Esto significa que no existe conexión entre el tamaño de la puntuación verdadera y el de los errores (puede haber calificaciones altas con errores bajos y bajas con errores altos).
- Tercer supuesto: los errores de medida de los sujetos de un test no están relacionados con los errores de medida de otro test diferente: $r(e_j, e_k) = 0$. Es decir, que los errores cometidos en un test no tienen que covariar de manera sistematica con los cometidos en otras ocasiones.

Ciertamente los supuestos desarrollados en relación con el modelo así como las deducciones que de ellos pueden hacerse no pueden ser comprobadas empíricamente, esto es, no permiten determinar cuál es la cantidad de error que presenta una puntuación concreta de un test. Por este motivo, Spearman introdujo la definición tests paralelos, esto es, aquellos tests que miden lo mismo exactamente, pero con diferentes ítems. Las puntuaciones verdaderas de los sujetos en los tests paralelos serían las mismas, y también serían iguales las varianzas de los errores de medida. De este modo resultaría posible

trabajar de forma empírica con las puntuaciones que las personas han obtenido en los tests.

A partir de la definición de este modelo, de los tres supuestos y de la definición de los tests paralelos, es posible llegar a fórmulas que permitan estimar la fiabilidad de los tests mediante dos estadísticos: el coeficiente de validez (la proporción de la varianza de las puntuaciones observadas aociada a la varianza de las puntuaciones verdaderas) y el error típico de medida (la desviación típica de las cantidades de error que afectan a las puntuaciones observadas).

Para estimar estos estadísticos es posible utilizar diversos procedimientos de recogida de datos derivados de la definición de tests paralelos (Prieto y Delgado, 2010)..

Para estimar empíricamente el error típico de la medida (ETM) (la precisión de las calificaciones de un sujeto, su variabilidad en torno a la puntuación verdadera), así como la fiabilidad (el cociente entre la varianza de las puntuaciones verdaderas y la de las puntuaciones observadas en una serie de sujetos) es posible utilizar diversos procedimientos de recogida de datos que reflejen distintas repeticiones del proceso de medida (Prieto y Delgado 2010), como:

- Test-retest: aplicación de un test a una muestra de sujetos en dos ocasiones entre las que el atributo se mantiene estable.
- Formas paralelas: aplicación a una muestra de sujetos en la misma ocasión o en ocasiones diferentes de dos versiones paralelas del test.
- Consistencia entre los apartados de una prueba: división del test en dos subconjuntos equivalentes de ítems o estimación a partir de las covarianzas entre los ítems de la prueba.
- Consistencia de las puntuaciones de distintos calificadores: evaluación de una muestra de conducta por calificadores independientes.

El análisis psicométrico de una prueba de ejecución o de desempeño mediante el modelo TCT suele realizarse sobre las puntuaciones a los atributos dadas por un calificador a cada candidato y sobre la puntuación total de los candidatos. Los estadísticos TCT empleados para analizar este tipo de pruebas son similares a los utilizados en las pruebas de selección múltiple (Prieto 2014): correlación ítem-test (discriminación de ítems), media y variabilidad de las puntuaciones de los candidatos en cada uno de los ítems y en su suma. Para obtener las estimaciones del coeficiente de fiabilidad se utiliza habitualmente el coeficiente alfa de Cronbach (1951) y adicionalmente, es posible obtener una estimación de la fiabilidad de las puntuaciones mediante la correlación entre las dos calificaciones independientes en la muestra total de candidatos.

5.2. La Teoría de la Generalizabilidad (TG)

Para determinar las fuentes del error de medida se formuló a mediados del pasado siglo una extensión de la TCT denominada Teoría de la Generalizabilidad (TG). La TG fue propuesta por Cronbach y su equipo de colaboradores (Cronbach, Rajaratnam y Gleser 1963; Gleser, Cronbach y Rajaratnam 1965 y Cronbach, Gleser, Nanda y Rajaratnam 1972) en la década de los 60 y primeros 70 (Martínez 1995, Martínez 2010, Muñiz 2010). Para enunciar dicha teoría, también conocida como TG, o teoría G, este grupo de autores se había basado en desarrollos teóricos de anteriores investigaciones (Burt 1936, Hoyt 1941, Ebel 1951 y Lindquist 1953) en las que se propugnaba el estudio de las propiedades de la fiabilidad de los tests mediante las técnicas estadísticas del análisis de la varianza (ANOVA). La TG igualmente se sirvió de los trabajos sobre las diferentes fuentes del error que habían realizado autores como el propio Cronbach (1947), Thorndike (1951) y Stanley (1971).

En el origen de la TG subyace la pretensión que esta teoría tenía de superar algunos de los problemas que planteaba la TCT. Sin embargo, a pesar de los evidentes avances, la TG no supuso una ruptura en relación con esta teoría, sino que puede considerarse una extensión de la misma. Entre los principales problemas de la TCT, destacamos: (i) una visión excesivamente restrictiva del concepto de paralelismo de las medidas; (ii) una concepción del error de la medida indiferenciada y unitaria, y (iii) unas limitaciones excesivas que la TCT impone a la fiabilidad y a la manera de realizar sus inferencias.

Con el fin de superar estos obstáculos, la TG propone las siguientes modificaciones a la TCT: (i) considera que las facetas son una muestra aleatoria de un universo mucho más amplio; (ii) utiliza el concepto estadístico de muestreo de fuentes de variación múltiple con la intención de considerar las diferentes características de la situación de medida como facetas del plan de medición, y (iii) sustituye los conceptos de puntuación verdadera y de fiabilidad por los de puntuación del universo y generalizabilidad, respectivamente (Mártínez, Hernández y Hernández 2006: 102).

En la TG existe una distinción conceptual entre los denominados Estudios de Generalizabilidad (Estudios-G) y los Estudios de Decisión (Estudios-D). Los primeros hacen referencia al planteamiento relativo al diseño del análisis para la obtención de los datos, a través de los cuales —por medio del ANOVA—, es posible estimar las varianzas de las diferentes facetas. Los Estudios-G son los que se realizan en la fase de construcción y desarrollo del instrumento de medida. En los segundos, los Estudios-D, gracias a los datos proporcionados por los estudios anteriores, se pueden resolver problemas concretos y tomar decisiones o extraer conclusiones en relación con las diferencias individuales o acerca de un determinado criterio (Cronbach, Gleser, Nanda y Rajaratnam 1972).

Tanto la TCT como la TG han desempeñado un papel relevante en el diseño de las evaluaciones y en la cuantificación de las fuentes de error. Sin embargo,

a pesar de sus evidentes ventajas como su sencillez matemática y su enjundia psicológica (Muñiz 2000) hay cuestiones que no es posible solucionar mediante el enfoque clásico. En el enfoque clásico las mediciones no son invariantes respecto al instrumento empleado, es decir, que las puntuaciones obtenidas por diversas personas en diferentes tests no están en la misma escala, cuando lo deseable científicamente sería que los resultados al emplear distintos instrumentos sí lo estuvieran. Por este motivo, los procedimientos de medición basados en la TCT no permiten determinar si las magnitudes de las calificaciones otorgadas por los examinadores se deben a que estos son excesivamente severos o benévolos o si la muestra calificada tiene alta o baja. Este hecho hace aconsejable utilizar modelos psicométricos que "permitan obtener la separabilidad de los parámetros de las personas y los calificadores (Tesio et al. 2015, citado por Prieto 2015b: 13).

Otra de las cuestiones que el enfoque clásico no ha resuelto convenientemente se refiere a la ausencia de invarianza de las propiedades de los tests respecto de los sujetos utilizados para esimarlos, lo que significa que propiedades psicométricas de los tests como la dificultad de los ítems o el índice de fiabilidad estarán en relación con el tipo de sujetos empleados para calcularlas. También existe evidencia empírica de que los tests no miden con la misma precisión a todos los sujetos, ya que esta precisión depende de nivel del sujeto en la variable medida. Las limitaciones de la TCT fueron descritas por Embretson y Hershberger en 1999.

El modelo TRI va a superar las limitaciones señaladas y va a generar una nueva tecnología psicométrica que permitirá complementar al modelo clásico expuesto. En concreto, el análisis por medio del modelo Rasch permite aportar información adicional a la TG, "as it is not limited to the

quantification of the different sources of error that affect observed scores. Being able to measure the elements of each facet using a common metric facilitates understanding of the different aspects that influence assessments and allows us to obtain measures of facet elements that are independent of the rest, and correct their idiosyncratic influence (2004)" (citado en Prieto y Nieto 2014: 373). Actualmente la TG se sigue empleando en la evaluación educativa para determinar la magnitud de las distintas fuentes de error que afectan a la varianza de las medidas.

5.3. La Teoría de la Respuesta al Ítem (TRI)

Es preciso esperar al desarrollo de los diferentes modelos de la TRI para observar un verdadero cambio en la manera de relacionar las medidas observadas con el constructo. Por medio de esta teoría es posible relacionar las puntuaciones de cada uno de los ítems con el constructo y estimar el valor de un sujeto en este constructo por medio de patrones de respuesta. Puede verse un análisis de las principales diferencias entre la TCT y la TRI y de las ventajas del modelo de Rasch en Andrich (1988), Bond y Fox (2001), Embretson y McCollam (2000), Embretson y Reise (2000), Hambleton, Swaminathan y Rogers (1991) y Wright y Stone (1979). Bock (1997) relata la historia de la TRI. En español, Muñiz y Hambleton (1992) realizan una revisión de las principales aportaciones de la TRI. También en español pueden consultarse Navas (1994), Barbero (1999), las tablas 5.1 y 5.2 de Mártínez, Hernández y Hernández (2006: 125 y s.). Podemos encontrar una síntesis útil en Muñiz (2010).

5.3.1. Modelos dicotómicos

5.3.1.1. Características de los modelos de la TRI

La TRI agrupa a una gran cantidad de diferentes modelos que, pese a sus diferencias, presentan una serie de rasgos básicos comunes (Mártínez, Hernández y Hernández 2006: 127-130). La TRI considera que la actuación de un sujeto en los ítems de un examen (comportamiento observable) es consecuencia de la existencia de un rasgo latente (por ejemplo, comprensión lectora), que no puede observarse directamente, pero que permite explicar o predecir la conducta de dicho sujeto. Esta relación entre la actuación de un sujeto en un ítem y el rasgo o conjunto de rasgos —que es necesario especificar— responsables de dicho rendimiento, se puede describir por medio de una función monótona creciente denominada *curva característica del ítem* (CCI) o *función de respuesta al ítem* (FRI).

En los modelos de la TRI es posible situar en un punto del espacio del rasgo o atributo —que en los modelos unidimensionales es una recta real—tanto ítems como personas. La probabilidad de que un sujeto responda correctamente a un ítem depende de la diferencia entre la capacidad del candidato y la dificultad del ítem. La colocación de las personas en el espacio del rasgo depende de la cantidad que tengan de este, mientras que los ítems se sitúan dependiendo de la cantidad de rasgo que exijan para su correcta ejecución. Este parámetro se denomina *dificultad* del ítem.

El *nivel de aptitud* asume en los modelos de la TRI el mismo papel predominante que en la TCT tenía el concepto de puntuación verdadera de un candidato. Este nivel de aptitud no depende, por tanto, del test concreto que estemos analizando ya que se trata de una variable continua y su distribución no precisa, por norma general, especificación. En consecuencia, el ítem se convierte en la unidad básica de medida en lugar del test y el

comportamiento o ejecución del individuo se determina a partir de las respuestas dadas a cada uno de los ítems.

5.3.1.2. Modelo de Rasch

El modelo de Rasch fue presentado en 1960 por el matemático danés Georg Rasch y se trata de uno de los modelos dicotómicos más conocidos de la TRI. Por medio de este modelo es posible representar el atributo objeto de la medición en una única dimensión en la que se sitúan conjuntamente personas e ítems (Rasch 1960/1980). Al conocer el nivel del candidato y la dificultad del ítem es posible determinar la probabilidad de que una respuesta sea correcta (Prieto y Delgado 2003), de tal forma que el nivel de aptitud de un candidato es independiente del test aplicado (Mártínez, Hernández y Hernández 2006: 130).

La ecuación que define al modelo Rasch, según Linacre (2012: 14), es la siguiente:

$$\log\left[\frac{P_{ni}}{1-P_{ni}}\right] = \mathrm{B}_n - \mathrm{D}_i, \qquad [1]$$

Donde,

- P_{ni} es la probabilidad de que la persona n supere con éxito el ítem i.
- B_n es la capacidad —nivel en la variable latente— de la persona *n*, y
- D_i es la dificultad del ítem *i*.

Según este modelo, la probabilidad de que una persona *n* responda correctamente el ítem *i*, depende de la diferencia entre la capacidad de la persona (B_n) y la dificultad del ítem (D_i). De modo que si la capacidad de la persona es igual a la dificultad del ítem, (es decir, $B_n = D_i$), la probabilidad de que la respuesta sea correcta es de .50.

El modelo de Rasch es capaz de situar en la misma métrica las diversas facetas que intervienen en el proceso de calificación y es muy utilizado actualmente

en evaluaciones educativas a gran escala como el programa PISA (OCDE 2006).

5.3.2. Modelos politómicos

Así como las respuestas a los ítems dicotómicos se codifican en dos categorías (acierto/error: 1/0), las respuestas a los ítems politómicos pueden clasificarse en más de dos categorías, tal y como puede suceder, por ejemplo, en los ítems de los tests de actuación. Para analizarlos se han desarrollado diversos modelos para TRI con ítems politómicos. Pueden consultarse al respecto: Boomsma, Van Duijn y Snijders (2001); Van der Linden y Hambleton (1997) y, especialmente, Ostini y Nering (2005). Presentamos a continuación dos modelos que son extensiones del modelo de Rasch: el Modelo de Crédito Parcial (PCM) y el Modelo de Escalas de Calificación (RSM) que es, a su vez, un caso particular del anterior.

5.3.2.1. Modelo de Escalas de Calificación

Los modelos de escalas de calificación (RSM, *Rating Scale Model*) (Andrich, 1978), añaden un parámetro umbral que representa el valor de paso entre dos valores adyacentes de una escala de calificación. Dicho valor, representado como F_k, es la posición en la que los valores adyacentes k y $k - 1$ tienen las mismas probabilidades de ser recibidos por el candidato; es decir, F_k representa el punto de transición en el que la probabilidad de que un candidato sea calificado con uno u otro de los dos valores adyacentes es de un 50% (cf. Bond y Fox 2007; Linacre 2006, 2010b y Andrich, 1998, 2005). En el modelo RSM se asume que los pasos entre los valores adyacentes son iguales entre las diversas facetas.

La ecuación que representa a este modelo es la siguiente (Eckes 2011: 11):

$$\mathrm{In}\left[\frac{P_{nik}}{P_{nik-1}}\right] = \mathrm{B}_n - \mathrm{D}_i - \mathrm{F}_k, \qquad [2]$$

Donde,

- p_{nik} es la probabilidad de que una persona *n* sea puntuada con la categoría *k* al ítem *i*,
- $p_{nik\text{-}1}$ es la probabilidad de que una persona *n* sea puntuada con la categoría *k* – 1 al ítem *i*,
- *k* es una categoría numérica de una escala de calificación que tiene *m* + 1 categorías ordinales,
- $\mathrm{F_k}$ es la dificultad de ser puntuado con la categoría *k* (en relación con k-1).

5.3.2.2. Modelo de Crédito Parcial

El Modelo de Crédito Parcial (PCM, *Partial Credit Model*) fue propuesto por Wright y Masters en 1982, por Masters en ese mismo año y por Masters y Wright en 1997, aunque es preciso subrayar que su primer desarrollo ya lo había realizado Andrich en su trabajo de 1978. Para una comparación entre el modelo RSM y el PCM ver Linacre (2000).

Al proceder de un modelo de un parámetro (1P) como el de Rasch, el PCM permite separar los parámetros de sujetos y de ítems. Además, al igual que sucede en el resto de modelos derivados del modelo de Rasch, es posible estimar el nivel de rasgo del candidato por medio de la puntuación total ya que se trata de un estadístico suficiente (Mártínez, Hernández y Hernández 2006: 208). El modelo PCM se utiliza habitualmente para analizar ítems que forman parte de tests de rendimiento educativo o ítems con respuestas en forma de escalas ordenadas, ya que permite otorgar calificaciones diferentes

—'crédito parcial'— a cada una de las partes que componen dichos ítems, en los que podría hablarse de valores de paso o dificultad que describirían las categorías ordenadas.

La ecuación que representa al modelo PCM es el siguiente (Eckes 2011: 12):

$$\mathrm{In}\left[\frac{P_{nik}}{P_{nik-1}}\right] = \mathrm{B}_n - \mathrm{D}_i - \mathrm{F}_{ik}, \qquad [3]$$

Donde,

- F_{ik} es la dificultad de recibir la categoría numérica *k* en el ítem *i* (en relación con k-1).

5.4. Descripción del modelo Many-Facet Rasch Measurement (MFRM)

Fue descrito teóricamente por John Michael Linacre en 1989 y dispone de un programa informático, denominado FACETS, que permite realizar los cálculos que describe el modelo (Linacre 2012). El modelo MFRM pertenece a una creciente familia de modelos de Rasch que, como indicábamos arriba, enlaza, a su vez, con los modelos comentados con anterioridad: el modelo RSM y el modelo PCM. Otros modelos son: el modelo lineal de prueba logística (LLTM, *Linear Logistic Test Model*) (Fischer 1973, 1995b; Kubinger 2009), el modelo mixto de Rasch (Rost 1990, 2004) y otros más (Eckes 2009). Puede consultarse una información detallada acera de los diferentes modelos Rasch en Fischer, 2007, Rost, 2001 y Wright y Mok, 2004.

El modelo MFRM es un procedimiento adecuado para analizar de forma simultánea diferentes facetas que pueden tener un impacto relevante en los resultados de evaluación (Eckes 2011: 12). Como en el resto de los modelos de Rasch, su principal característica es la *invariancia de la medida,* también

llamada *objetividad específica* (Bond y Fox 2007; Engelhard 2008; Fischer 1995a), lo que significa que las medidas de los candidatos son independientes del resto de las facetas (tareas, calificadores...). Esta objetividad específica es una propiedad de los modelos de Rasch que no poseen otros modelos TRI como los derivados de los modelos de dos y tres parámetros de Birnbaum (1968). Otra propiedad específica de los modelos de Rasch es la *suficiencia*, que implica que la puntuación bruta obtenida por un examinado es el estadístico suficiente para estimar su parámetro en la escala logit. Asimismo, la suma de las calificaciones otrogadas por un calificador a un grupo de candidatos es el estadístico suficiente para estimar el parámetro del calificador en la escala de severidad. De este modo, y por medio del modelo, es posible obtener de manera independiente estimaciones en una misma escala de los diferentes parámetros de cada una de las facetas implicadas en la evaluación, que son las que pueden contribuir a la variabilidad de las medidas. Además, es posible representar "la contribución aditiva de cada faceta al *logit* o logaritmo del cociente entre la probabilidad de que una persona reciba una calificación en una tarea (por ejemplo, 3) y la probabilidad de que reciba la calificación inmediatamente inferior (2)" (Prieto 2011: 234).

Desde su primera definición teórica, el modelo MFRM se ha usado en diversos ámbitos como la evaluación de lenguas, la medición en educación y psicología, las ciencias de la salud y muchos otros. Puede consultarse a este respecto Bond y Fox 2007; Engelhard 2002; Harasym, Woloschuk y Cunning 2008, McNamara 1996; y Wolfe y Dobria 2008. Un ejemplo de la importancia que este modelo tiene en el aprendizaje, la enseñanza y la evaluación de lenguas en el ámbito europeo es la utilización que se hizo de él para el desarrollo de los descriptores ilustrativos de las escalas de medición del dominio de la lengua durante el proceso de gestación del *Marco de referencia*: 205; también se empleó el *Marco de referencia* en la realización de los DVD (Consejo de Europa 2008) con ejemplos de producciones orales de jóvenes de entre 13 y 18 años que ilustran los 6 niveles del *Marco de referencia* que

patrocinó la *División de Política Lingüística* del Consejo de Europa (Corrigan 2007) y en la que colaboraron Cambridge ESOL, el Instituto Cervantes, la Fundación Eurocentres, el Instituto Goethe y el Centro de evaluación y de certificación de la Universidad de Perugia. El propio Bryan North, uno de los autores del *Marco de referencia*, reconocía (1996/2000: 349) lo extraordinariamente pertinente que había resultado este modelo para el desarrollo del *Marco de referencia* (Eckes 2011: 13, North 2008, North y Jones 2009, y North y Schneider 1998). También el modelo MFRM constituye una de las secciones —en concreto la H, redactada por Thomas Eckes—del suplemento (Consejo de Europa 2009b) que acompaña al *Manual para relacionar exámenes con el* Marco de referencia (2009a).

En la evaluación de las destrezas productivas, se ha aplicado este modelo en casos en los que un evaluador califica la actuación de un candidato en una tarea por medio de un conjunto de categorías ordenadas. En concreto, se ha utilizado MFRM para analizar las facetas intervinientes en exámenes de expresión oral y escrita de inglés como segunda lengua (Sudweeks, Reeve y Bradshaw 2005, Park 2004; Tyndall y Kenyon 1996), de japonés (Kondo-Brown 2002), de alemán (Eckes 2004, 2005, 2008, 2010, 2011), de francés (Casanova y Demeuse 2011) y de español (Prieto 2011, Prieto y Nieto 2014).

Con el fin de adaptar los modelos que presentaremos seguidamente al caso que nos ocupa, en adelante, a las *personas* las denominaremos *candidatos* (el DELE es un título oficial que acredita que los candidatos que lo superan tienen el grado de competencia y dominio del español que se exige en cada uno de los seis niveles) y a los *ítems tareas* (la prueba de EIE del nivel A2 está compuesta por tres tareas).

5.4.1. Análisis de las facetas

Cuando en un proceso de evaluación los calificadores evalúan un único atributo del desempeño del candidato en una tarea es posible analizar, al

menos, dos facetas: calificadores y candidatos. En caso de que los candidatos se enfrenten a más de una tarea y los calificadores, en consecuencia, califiquen su actuación en las mismas de manera independiente, es necesario tener en cuenta una tercera faceta: la tarea. Si las tres facetas están claramente definidas, la expresión formal del modelo MFRM sería la siguiente (Eckes 2011: 14):

$$\mathrm{In}\left[\frac{P_{nljk}}{P_{nljk-1}}\right] = \mathrm{B}_n - \mathrm{D}_l - \mathrm{R}_j - \mathrm{F}_k, \qquad [4]$$

Donde,

- P_{nljk} es la probabilidad de que un candidato *n* reciba la calificación *k* en la tarea *l* por el calificador *j*.
- $P_{nljk\text{-}1}$ es la probabilidad de que un candidato *n* reciba la calificación inferior *(k-1)* en la tarea *l* por el calificador *j*.
- B_n es la competencia —valor de la variable latente— del candidato *n*.
- D_l es la dificultad de la tarea *l*.
- R_j es la severidad del calificador *j*.
- F_k es la dificultad de recibir la calificación de *k* en relación con la categoría adyacente inferior *k*-1.

En caso de que la segunda faceta en cuestión sea *criterios de calificación* en lugar de *tareas* y, siempre que se asuma una estructura constante de la escala de calificación en relación con los elementos de las facetas, el modelo quedaría de este modo (Eckes 2011: 19):

$$\mathrm{In}\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \mathrm{B}_n - \mathrm{C}_i - \mathrm{R}_j - \mathrm{F}_k, \qquad [5]$$

Donde,

- C_i = dificultad del criterio de calificación *i*.

Como se puede observar en las ecuaciones anteriores, el modelo MFRM se caracteriza por ser un modelo lineal aditivo, basado en la transformación logística de las calificaciones observadas en una escala de *logit*. En esta ecuación, la variable dependiente es el logaritmo natural de cociente $\frac{P_{nijk}}{P_{nlik-1}}$, mientras que las diversas facetas son las variables independientes que influyen en este registro de probabilidades.

El término F_k indica cómo se deben utilizar los datos. En concreto, en las ecuaciones 4 y 5, este parámetro especifica que se debe usar un modelo RSM en todos los elementos de cada faceta.
Si las facetas fueran cuatro en lugar de tres, el modelo sería:

$$\mathrm{In}\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \mathrm{B}_n - \mathrm{D}_i - \mathrm{R}_j - \mathrm{C}_l - \mathrm{F}_k, \qquad [6]$$

5.4.2. Modelos híbridos

La aplicación del modelo RSM a las facetas supone que los pasos entre los diversos valores adyacentes son idénticos para las diferentes facetas, es decir, que todas las facetas a las que se les aplica el modelo comparten idéntica estructura. Esto significa, por ejemplo, que una puntuación en un criterio de calificación concreto —un 2 en Adecuación— se supone que es funcionalmente equivalente a la misma puntuación en el resto de criterios —2 en Corrección y alcance — (Eckes 2011: 90) lo que significa que los valores de paso son constantes en todos los criterios de calificación. Lo mismo sucede cuando se aplica este modelo a la faceta de calificadores, que únicamente se obtiene un resumen de cómo el equipo de calificadores, en su conjunto, utiliza cada una de las bandas de calificación (que corresponden a un descriptor ilustrativo con el que el calificador compara la actuación del

candidato) de las diferentes escalas de criterios. El análisis de las tareas con el modelo RSM supone que no es posible determinar si los atributos son utilizadas de manera uniforme por los calificadores en todas las tareas.

Cuando no se supone que los valores de paso son constantes en cada una de las variables independientes o facetas que intervienen en el proceso es necesario utilizar el modelo PCM. De este modo, diversos autores han propuesto modelos híbridos derivados de las ecuaciones anteriores (vid. Linacre 1989; Linacre y Wright 2002: 486-487; Myford y Wolfe 2004b: 501-506 y Eckes 2011: 90-95) que combinan ambos modelos o utilizan únicamente el modelo PCM.

5.4.2.1. Modelo híbrido número 1

Con el fin de localizar la variable de paso en cada uno de los criterios de calificación y determinar, de este modo, si a los candidatos les ha resultado igual de fácil o complicado alcanzar la misma calificación en cada uno de ellos, es necesario modificar en la ecuación 5 la variable independiente F_k y añadirle un doble subíndice: F_{ik}.

El modelo revisado queda del siguiente modo:

$$\mathrm{In}\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \mathrm{B}_n - \mathrm{D}_i - \mathrm{R}_j - \mathrm{F}_{ik}, \qquad [7]$$

en el que todos los parámetros son idénticos a los ya reseñados, excepto el término F_{ik}, que expresa la localización del valor de paso entre los valores adyacentes *k* y *k* – 1, en el criterio *i*. Con esta modificación, FACETS proporciona información acerca de cómo el conjunto de calificadores utiliza las escalas de calificación de cada uno de los criterios.

Se trata, en consecuencia, de un modelo híbrido que combina un modelo PCM aplicado a los criterios de calificación con un modelo RSM aplicado a los calificadores (Myford y Wolfe 2004a, 2004b).

5.4.2.2. Modelo híbrido número 2

Si el objetivo es analizar cómo aplica cada uno de los calificadores el valor de paso entre *k* y *k* – 1 en el conjunto de criterios de calificación, es necesario modificar la variable independiente F_k de la ecuación 5, aunque de un modo diferente a como habíamos hecho en la anterior —la siete—. Con este fin añadimos el doble subíndice: F_{jk}.

El modelo revisado queda del siguiente modo:

$$\mathrm{In}\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \mathrm{B}_n - \mathrm{D}_i - \mathrm{C}_j - \mathrm{F}_{jk}, \qquad [8]$$

en el que, al igual que sucedía con la ecuación 7, todos los parámetros son idénticos a los reseñados en la ecuación 5, excepto el término F_{jk}, que expresa la localización del valor de paso entre los valores adyacentes *k* y *k* – 1 para cada uno de los calificadores *j*. Con esta modificación, FACETS proporciona información acerca de cómo cada uno de los calificadores utilizan las escalas de calificación en el conjunto de criterios (Linacre y Wright 2002: 487).

Se trata de un modelo híbrido que combina un modelo PCM aplicado a los calificadores con un modelo RSM aplicado a los criterios (Myford y Wolfe 2004a, 2004b).

También resulta factible incorporar una nueva faceta a este modelo híbrido. El modelo revisado queda del siguiente modo:

$$\mathrm{In}\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \mathrm{B}_n - \mathrm{D}_i - \mathrm{R}_j - \mathrm{C}_l - \mathrm{F}_{jk}, \qquad [9]$$

5.4.2.3. Modelo híbrido número 3

También es posible plantear modelos en los que se combina la acción conjunta de varias facetas. Si el objetivo del análisis es estudiar cómo utiliza cada calificador cada uno de los descriptores de las escalas de calificación de cada criterio de calificación es necesario modificar la variable independiente F_k de la ecuación 5, aunque de un modo diferente a como hemos hecho en los dos modelos híbridos anteriores. En este supuesto se añade un triple subíndice: F_{ijk}.

El modelo revisado queda del siguiente modo:

$$\mathrm{In}\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \mathrm{B}_n - \mathrm{D}_i - \mathrm{R}_j - \mathrm{F}_{ijk}, \qquad [10]$$

en el que, al igual que sucedía con los dos modelos anteriores, todos los parámetros son idénticos a los reseñados en la ecuación de referencia, excepto el término F_{ijk}, que expresa la localización del valor de paso entre los valores adyacentes *k* y *k* – 1 en el criterio de calificación *i* para cada el calificador *j*. Con esta modificación, FACETS proporciona información acerca de cómo cada uno de los calificadores utiliza las escalas de calificación en cada uno de los atributos (Linacre y Wright 2002: 487).

Se trata de un modelo híbrido que combina dos Modelos de Crédito Parcial, uno aplicado a los criterios de calificación y otro aplicado a los calificadores (Myford y Wolfe 2004a, 2004b).

5.4.3. Características del modelo MFRM

Para realizar los análisis con el modelo MFRM hemos utilizado el programa *FACETS*, en concreto la versión 3.70.0 (Linacre 2012). FACETS permite estimar "los parámetros mediante el método de estimación conjunta por máxima verosimilitud (JML)" (Prieto 2011: 234).

El modelo MFRM permite estimar de manera diferenciada e independiente los parámetros de las diversas facetas en una escala común y, como consecuencia de la combinación con el modelo PCM, se asume que pueden variar los pasos entre las categorías k y k-1 en las diversas facetas. La suma de todas las puntuaciones otorgadas a la ejecución de los examinados en una tarea por todos los calificadores es el estadístico suficiente para estimar el *logit* de la tarea; la suma de todas las calificaciones otrorgadas por una calificador es el estadístico suficiente para estimar el *logit* del calificador; la suma de todas las puntuaciones "holísticas" de todos los examinados en todas las tareas es el estadístico suficiente para estimar el *logit* del atributo "Calificación Holística", etc. (Linacre y Wright 2002).

La escala *logit* puede oscilar entre 0, que convencionalmente se fija en el nivel medio de las facetas (tareas, calificadores, atributos…), y $\pm \infty$; la posición de los candidatos puede variar libremente en dicha escala. En el caso de los calificadores, los valores superiores a 0 indican mayor severidad y los inferiores a 0 menor severidad o mayor benevolencia. Como se expondrá en detalle más adelante, los valores superiores o inferiores a 0 indican mayor o menor competencia de los candidatos. En la graduación de las tareas y los atributos, los valores superiores o inferiores a 0 indican mayor o menor dificultad.

Además de estimar el nivel en *logit* de cada calificador, el modelo permite obtener un error típico de la medida (SE= precisión del valor estimado) e índices de ajuste entre las calificaciones observadas y las predichas por el

modelo.

Las predicciones del modelo sobre las calificaciones de un examinador se derivan de los siguientes supuestos (Engelhard 2013):

1. *Unidimensionalidad*: los elementos de las facetas se localizan simultáneamente en una única variable (el mapa de la variable).
2. *Calibración independiente de los calificadores*: la calibración de los calificadores en el mapa de la variable es independiente de los candidatos utilizados en la calibración.
3. Cualquier candidato debe tener mayor probabilidad de obtener una calificación más alta de los calificadores benévolos que de los severos.

El análisis facilita estadísticos a nivel individual y grupal. Los estadísticos a nivel individual son una medida en *logit* para cada elemento de cada faceta, un error típico de medida (SE= precisión del valor estimado) e índices de ajuste entre las respuestas observadas y las predichas por el modelo. A nivel grupal son (Myford y Wolfe 2004a, 2004b; Prieto 2011: 234) el ajuste promedio, la media, la variabilidad y la fiabilidad de las medidas de las personas, las tareas y los calificadores.

5.4.4. Estadísticos básicos

Los estadísticos básicos del modelo MFRM son:

- *Estadísticos descriptivos de la escala de severidad*. Se puntúa en la escala *logit* el grado de severidad/benevolencia de los calificadores de acuerdo con el modelo. Convencionalmente el punto 0 se sitúa en el valor medio de la severidad de los calificadores que participan en el proceso de calificación. Los valores superiores a 0 indican mayor severidad y los inferiores menor severidad (o mayor

benevolencia). La localización del punto 0 puede variar entre distintos exámenes, por lo que la escala no es absoluta: el aspecto que se valora es la diferencia en severidad de cada calificador respecto a la media de severidad en el examen (0). En cada prueba de EIE de un examen se utiliza la desviación típica de los valores de severidad de los calificadores para cuantificar la magnitud de su variabilidad en severidad.

La significación estadística de la variabilidad de los calificadores se contrasta por medio del estadístico chi-cuadrado. Este estadístico permite contrastar la hipótesis de que todos los calificadores ejercen el mismo nivel de severidad al calificar las tareas de los candidatos. Un chi-cuadrado significativo ($p<,05$) indica que dicha hipótesis puede ser rechazada (al menos dos calificadores presentan diferencias fiables en severidad).

El promedio observado (*Observed Average*) se obtiene sumando las calificaciones asignadas por el examinador y dividiendo por el número de calificaciones. El problema es que el promedio observado depende tanto de la severidad del calificador como de la competencia de los candidatos. Un calificador puede presentar un promedio mayor que otro, bien porque haya sido más benévolo calificando, bien porque los candidatos que ha calificado tenían mayor grado de competencia. Por ello, si se desea un indicador de la severidad de los calificadores en la escala de puntuaciones brutas es necesario recurrir a al promedio imparcial (*Fair Average*) (Eckes 2011: 60). En cada prueba de EIE la media imparcial puede oscilar entre 0 y 3: los valores altos revelan menor severidad que los bajos (Prieto 2015a). El promedio observado para el calificador j, esto es, $\mathrm{M}_{O(j)}$, es la calificación media del calificador a todos los candidatos en los atributos utilizados en la calificación:

$$\mathrm{M}_{O(j)} = \frac{\sum_{n=1}^{N} \sum_{i=1}^{I} X_{nij}}{N \, . \, I} \qquad [11]$$

Un problema no trivial con promedios observados es que confunden la severidad del calificador y la competencia del candidato.

Cuando las tres facetas están claramente definidas, la ecuación anterior (11) se convierte en (Eckes 2011: 61):

$$\mathrm{In}\left[\frac{P_{jk}}{P_{jk-1}}\right] = \mathrm{B}_M - \mathrm{D}_M - \mathrm{R}_j - \mathrm{F}_k, \qquad [12]$$

Donde P_{jk} es la probabilidad de que el calificador *j* utilice la puntuackión *k* al evaluar a los candidatos en los atributos, B_M es la media de competencia de los candidatos y D_M la media de dificultad de los atributos.

La fórmula de cálculo del promedio imparcial (*Fair Average*) para el calificador j, esto es, $\mathrm{M}_{F(j)}$, es la siguiente:

$$\mathrm{M}_{F(j)} = \sum_{r=0}^{m} rp_{jr} \cdot \qquad [13]$$

- *Índices de ajuste*. Indican el grado en el que las calificaciones observadas (las otorgadas por los calificadores a los candidatos en las tareas) se diferencian de las esperadas (las predichas por el modelo, según el nivel del candidato, la severidad del calificador y la dificultad de la tarea). Los índices de ajuste se denominan *Infit* y *Outfit* y son las medias de los cuadrados de las diferencias

estandarizadas. *Outfit* es la media no ponderada de dichos valores, por lo que es muy sensible a desajustes extremos, mientras que *Infit* es la media de los valores ponderados con la función de información (Wolfe 2009). Ambos estadísticos tienen un valor esperado de 1, aunque su valor puede oscilar entre 0 e ∞. Los valores menores que 1 evidencian que los residuos (diferencias entre los valores observados y esperados) son menores que los esperados por azar y se puede interpretar como sobreajuste. Cuando los valores son superiores a 1 significa que el desajuste es superior al esperado. Linacre (2012) indica que un desajuste superior a 2 se considera desajuste severo que degrada las medidas. FACETS aporta valores individuales de ajuste para candidatos, calificadores, ítems y atributos.

- *Correlación calificador-resto de calificadores, (SR/ROR, correlation).* Cuantifica el grado de consistencia de las calificaciones de cada uno de los calificadores con las del resto. Convencionalmente, se considera que los valores inferiores a 0,30 son demasiado bajos y corresponden a calificadores inconsistentes que realizan una ordenación de candidatos diferente al resto.

- *Separabilidad de las medidas* (*SR, separation reliability*). FACETS, además de evaluar la precisión individual de las medidas (de cada persona, cada calificador o cada tarea), proporciona una cuantificación del grado de separabilidad de las medidas de los elementos de cada faceta. La separabilidad indica en qué grado la variabilidad se debe a la variabilidad verdadera (sin error de medida): la proporción de la varianza verdadera respecto a la varianza observada de las medidas. Las interpretaciones sustantivas de este índice son ligeramente diferentes, según se trate de una faceta u otra (Myford y Wolfe 2004b: 500):

- *La fiabilidad de la separación de las medidas de los candidatos* (*PSR, Person separation reliability*) es un índice comparable al coeficiente alfa de la TCT e indica cuál es la proporción de la varianza verdadera respecto de la varianza observada de los candidatos calificados (Wright y Master 1982):

  $$PSR = 1 - ((\text{Media } (SE^2_{Bn}) / \text{Varianza } (B_n)), \quad [14]$$

  Donde,

  - SE_{Bn} es el error típico de medida del valor del sujeto *n* en la variable.

  Cuando las medidas reflejan con fiabilidad la variabilidad de los candidatos en el constructo es esperable que los valores de PSR sean altos.

- *La fiabilidad de la separación de las medidas de los calificadores* (*RSR, Rater separation reliability*) refleja variaciones sustanciales en el nivel de severidad entre los calificadores. Si los calificadores utilizan las escalas de calificación de manera similar es esperable que el valor sea bajo.

- *La fiabilidad de la separación de las medidas de los ítems o criterios de calificación* (*ISR, Item separation reliability*) hace referencia a la fiabilidad de las estimaciones de la dificultad de los ítems con los que se mide un atributo. Se esperan valores altos de ISR cuando se incluyen criterios con diferente dificultad con el fin de garantizar un adecuado muestreo de los distintos valores del constructo evaluado (Wright y Stone 1979; Prieto 2011).

- *Estadísticos de las categorías de evaluación*. Se pueden utilizar diferentes indicadores para determinar si las categorías son funcionales empíricamente, es decir, si están ordenadas y son distinguibles:

  - *Orden de los promedios en las categorías de las medidas de las personas*. Los promedios de las medidas (*logit*) de las personas que reciben una calificación deben estar ordenados monotónicamente si funcionan adecuadamente las categorías de evaluación. Según revela este patrón de resultados, cuanto mayor sea la calificación recibida, mayor será el nivel de las personas en el constructo (Park 2004).

  - *Outfit*. FACETS calcula para cada atributo la media de los cuadrados de los residuos estandarizados (diferencias entre los valores observados y predichos por el modelo). Si las diferencias son muy pequeñas, Outfit adoptará un valor próximo a 1,0. Los valores de Outfit superiores a 2,0 indican que la categoría de evaluación no ha sido utilizada de manera adecuada.

  - *Orden de los pasos entre las categorías* (Linacre 2002a). Es preciso observar si los pasos entre las categorías están ordenados monotonicamente y si se encuentran suficientemente separados. Si estos pasos no se encuentran ordenados significa que existen categorías que no son las de más probable uso en ningún rango de la variable medida (Prieto 2011: 235).

CAPÍTULO 6

Método

6.1. Participantes

En la convocatoria de mayo de 2012 se presentaron 4301 candidatos al examen para la obtención del Diploma de Español Nivel A2, uno de los certificados del sistema de certificación DELE. Para calificar las muestras de la prueba de EIE se utilizaron dos procedimientos diferentes, tal y como se expondrá a continuación en el § 6.4. *Procedimiento*. El 88,7% de los candidatos (3858) se calificaron por medio del sistema de doble calificación: se formaron parejas de calificadores y a cada pareja se le asignaron los mismos exámenes. Con el fin de que la calificación se realizara de manera independiente, uno de los calificadores calificó los exámenes originales y el otro fotocopias de los mismos.

Para la calificación del restante 10,3% de candidatos (443) se estableció una red entre los doce calificadores que participaron en el proceso de calificación, de forma que todos compartieron pruebas con el resto de examinadores. El proceso de calificación se realizó mediante un sistema informático de visualización automatizada de pruebas denominado Hares (Instituto Cervantes s.f. a).

Los candidatos que se presentaron en la convocatoria de mayo de 2012 realizaron, según el centro de examen en el que presentaban, uno de los dos exámenes que tuvieron lugar en fechas diferentes: el 25 (Instituto Cervantes 2012a) y el 26 de mayo (Instituto Cervantes 2012b). De los 443 candidatos que fueron calificados mediante este sistema de calificación en red, 259 se

presentaron al examen del día 25 (el examen se codificó con el nombre D031) y 184 al del 26 (el examen se codificó con el nombre D032).

En la *Guía del examen del Diploma de español Nivel A2* (Instituto Cervantes 2013) se informa de que:

> El Diploma de español nivel A2 acredita la capacidad del usuario de la lengua para comprender y utilizar expresiones cotidianas de uso frecuente, relacionadas casi siempre con áreas de experiencia que le sean especialmente relevantes por su inmediatez (información básica sobre sí mismo y sobre su familia, compras y lugares de interés, ocupaciones, etc.); para realizar intercambios comunicativos sencillos y directos sobre aspectos conocidos o habituales y para describir en términos sencillos aspectos de su pasado y de su entorno, así como para satisfacer cuestiones relacionadas con sus necesidades inmediatas.

Por medio de las pruebas y tareas que conforman el examen Diploma de Español Nivel A2 se evalúan conocimientos y destrezas en diversas actividades comunicativas de la lengua: comprensión, expresión e interacción, contextualizadas en los cuatro ámbitos que define el *Marco de referencia*: personal, público, educativo y profesional.

El repertorio lingüístico que describe el nivel de competencia lingüística que debe ser capaz de manejar, de forma productiva y receptiva, el candidato que aspira a superar este nivel es el que figura en los inventarios de material lingüístico del *Plan curricular*, documento desarrollado por el Instituto Cervantes a partir de las escalas de descriptores del *Marco de referencia*. En el *Plan curricular* se desglosan las competencias comunicativas de la lengua que recogen los contenidos del examen en los siguientes componentes e inventarios: Componente gramatical (Gramática, Pronunciación y prosodia, Ortografía), Componente pragmático-discursivo (Funciones, Tácticas y estrategias pragmáticas, Géneros discursivos y productos textuales) y Componente nocional (Nociones generales, Nociones específicas).

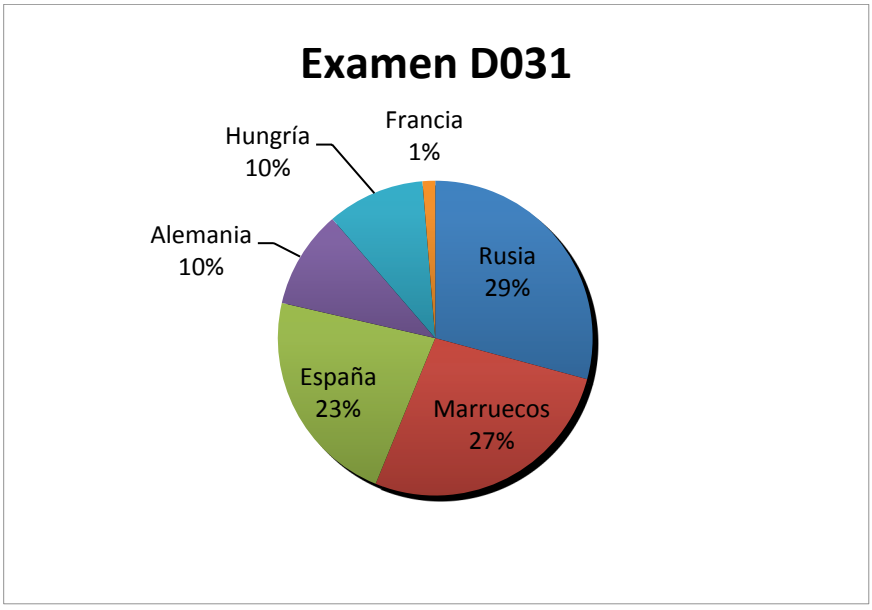
6.2. Norma lingüística

En el Diploma de Español Nivel A2 se emplean textos de entrada (orales y escritos) procedentes de diversas fuentes y variedades del español peninsular contemporáneo. En los textos producidos por el candidato se considera válida cualquier norma lingüística que se siga coherentemente y que esté respaldada por grupos amplios de hablantes cultos.

La procedencia geográfica de los candidatos que fueron calificados fue la siguiente:

- Examen D031 (gráfico 1): Brasil (0,39%)[7], Francia (1,54%), Alemania (11,97%), Hungría (1,93%), Marruecos (32,05%), Rusia (34,71%) y España (26,64%).

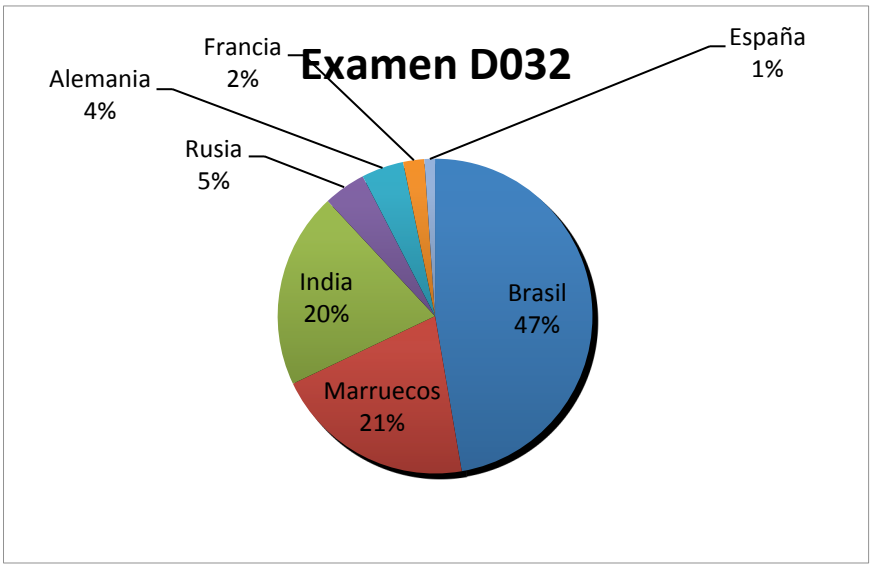
Gráfico 1
Procedencia de los alumnos presentados



[7] Brasil no figura en el gráfico por no alcanzar el 1%.

- Examen D032 (gráfico 2): Brasil (47,28%), Francia (2,17%), Alemania (4,34%), India (20,11), Marruecos (20,65%), Rusia (4,35%) y España (1,09%).

Gráfico 2
Procedencia de los alumnos presentados



6.3. Instrumento

La prueba de EIE del Diploma de Español Nivel A2 se presenta en un único cuadernillo en el que aparecen las tareas y en el que se deben redactar las respuestas. La prueba consta de tres tareas: dos de interacción y una de expresión. La duración total de la prueba es de 50 minutos. La extensión total de palabras que los candidatos deben escribir en el espacio reservado para cada tarea entre los tres textos oscila entre 170 y 200 palabras.

La tarea número 1 consiste en completar campos abiertos de formularios, encuestas, fichas…, o espacios para comentarios en foros virtuales (blogs, noticias, etc.), en los que el candidato debe escribir entre 30-40 palabras. La

información introducida por el candidato debe estar basada en el texto de entrada.

El objetivo de esta tarea es evaluar la capacidad del candidato para intercambiar información personal sencilla, relacionada con su entorno próximo o con asuntos cotidianos como hábitos, deseos, gustos, preferencias, etc.; intercambiar información breve y sencilla sobre asuntos prácticos de la vida cotidiana (confirmaciones, reservas, citas, disculpas, etc.); escribir oraciones breves y básicas para hacer descripciones, por medio de listas de elementos muy sencillos, sobre personas, objetos, lugares o acciones.

Los tipos de textos que se utilizan en esta primera tarea son impresos, solicitudes o espacios para comentarios en foros, noticias, etc., con apoyo gráfico o escrito, que explicitan las pautas que el candidato debe seguir.

En concreto, en la forma de examen del día 25 de mayo de 2012 (Instituto Cervantes 2012a), se le pidió al candidato que redactara un anuncio en una página web de una escuela de idiomas en el que presentara e indicara qué lengua quiere practicar, informara de cuándo quiere hacer el intercambio y propusiera un lugar y una hora para encontrarse.

En la forma de examen del día 26 de mayo de 2012 (Instituto Cervantes 2012b), se le pidió al candidato que redactara un anuncio en la sección de empleos de un periódico digital para buscar una persona que cuidara su casa. Debía explicar por qué necesitaba dicha persona, describir dónde vivía y cómo era la casa e informar del horario y del precio.

En la segunda tarea el candidato debía redactar un texto epistolar de entre 70 y 80 palabras del ámbito personal: carta breve, correo electrónico, nota, mensaje o postal.

El objetivo de esta tarea es evaluar la capacidad del candidato para intercambiar por escrito, a través de cartas, correos electrónicos, notas,

mensajes o postales, información de su entorno personal, relacionada con asuntos cotidianos: actividad académica o profesional actual o del pasado, lugar de trabajo o estudios, residencia temporal o permanente, condiciones de vida, estado general de las cosas, experiencias de estudio o de trabajo, etc.

Los tipos de textos que el candidato debe redactar y presentar en el formato del género al que pertenece en esta segunda tarea son cartas, correos electrónicos, notas, mensajes o postales, con apoyo escrito, que son las pautas que el candidato debe seguir.

En la forma de examen del día 25 se le pidió al candidato que escribiera un correo electrónico a un/a amigo/a en el que le informara de que se había cambiado de casa y describiera dónde estaba la casa y cómo era, explicara los motivos por los que se había cambiado de vivienda e invitara a su amigo/a y le explicara cómo llegar. El correo electrónico debía constar de saludo y despedida.

En la forma del día 26 se le pidió al candidato que escribiera un correo electrónico a un/a amigo/a en el que le informara de a quién había visto, cuándo y dónde, explicara con quién había ido y por qué y contara por qué le gustaba ese grupo/cantante y desde cuando. El correo electrónico debía constar de saludo y despedida.

En la tercera tarea el candidato debe redactar una composición breve, de entre 70 y 80 palabras: descripción, narración, entrada en un diario o biografía a partir de los datos que se proporcionan.

El objetivo de esta tarea es evaluar la capacidad del candidato para redactar textos descriptivos o narrativos.

Se proporciona al candidato información clara y sencilla relativa al contenido del texto que tiene que escribir por medio de diagramas, mapas, tablas, gráficos, iconos, etc., con la finalidad de que ayude al candidato a acotar y

contextualizar el texto.

En la forma de examen del día 25 se le pidió al candidato que escribiera una redacción sobre la persona más importante en su vida, en la que comentara cómo era (aspecto físico y personalidad), cuáles eran sus aficiones y por qué era importante (qué es lo que más le gustaba de ella), cuándo había sido la última vez que la había visto y en qué lugar y qué le gustaría hacer con ella.

En la forma de examen del día 26 se le pidió al candidato que escribiera una biografía de la vida de su profesora de español en la que comentara cómo era de niña y cómo era en aquel momento (aspecto físico y personalidad), dónde vivía, con quién y dónde vivía en aquel momento y qué hace habitualmente y qué cosas le gustan. Se le proporcionan al candidato algunos breves datos biográficos de la profesora como ayuda.

6.4. Procedimiento

En la evaluación de las pruebas de EIE de los dos exámenes participaron doce calificadores. Todos los examinadores calificaron tareas de ambas pruebas. La manera de organizar la distribución de las pruebas de los candidatos entre los calificadores tiene una gran relevancia en nuestro trabajo. Teóricamente, si no existieran limitaciones de tiempo ni de horas de trabajo, lo ideal sería que todos los examinadores calificaran a todos los candidatos. De este modo, al hallar la media aritmética de las puntuaciones de los calificadores, se sabría cuál es su grado de severidad/benignidad. Linacre y Wright (2002) dan noticia de un plan de calificación realizado por la organización estadounidense College Board (Engelhard y Myford 2003, Braun 1988). Reproducimos la figura en la tabla 5.

Tabla 5

Judge Essay	1 ABC	2 ABC	3 ABC	4 ABC	5 ABC	6 ABC	7 ABC	8 ABC	9 ABC	10 ABC	11 ABC	12 ABC
Person1	553	686	877	687	777	685	565	667	586	567	776	696
2	454	542	445	534	334	344	433	526	444	445	533	534
3	434	544	343	555	433	544	563	443	554	454	443	343
4	345	426	232	545	445	225	464	456	642	446	445	335
5	443	548	656	545	657	448	558	466	464	448	547	348
6	544	846	843	565	633	367	788	673	666	566	564	454
7	545	665	454	667	755	646	773	785	874	565	745	447
8	553	763	655	675	775	653	773	656	784	576	573	574
9	343	643	643	645	534	523	665	674	753	546	545	765
10	564	766	884	776	655	667	875	778	778	667	649	888
11	535	524	537	544	545	435	546	557	326	446	456	334
12	436	644	444	546	666	555	574	445	745	356	763	676
13	445	486	657	566	246	366	368	448	467	348	569	349
14	446	533	333	344	545	343	463	353	354	346	462	363
15	548	855	743	746	766	656	665	765	854	666	862	844
16	644	653	547	545	643	454	556	467	666	447	558	667
17	414	817	625	628	536	518	425	618	717	627	639	436
18	334	655	443	445	243	473	445	747	654	445	435	334
19	747	745	837	756	755	847	664	688	737	656	847	938
20	443	666	735	556	557	557	588	667	666	557	476	488
21	242	443	336	465	245	243	263	245	441	253	342	254
22	564	765	747	666	864	577	667	576	667	557	667	785
23	446	566	753	646	444	565	475	388	576	557	557	776
24	332	422	334	433	322	214	423	223	323	313	233	223
25	543	664	544	657	646	544	454	448	547	545	456	464
26	644	764	955	756	545	658	655	867	776	646	756	885
27	342	346	334	344	346	234	256	256	345	345	256	253
28	343	463	335	334	465	573	341	475	442	243	462	272
29	433	444	323	446	334	333	235	336	423	336	323	343
30	542	564	244	655	445	224	546	575	645	446	432	555
31	325	514	313	425	315	314	334	225	525	314	324	314
32	644	744	445	545	533	553	567	584	664	447	556	364

Figura 1 de Linacre y Wright (2002)

Lamentablemente, cuando el número de candidatos es elevado y el número de calificadores limitado esta propuesta resulta inviable.

Con el fin de soslayar este impedimento se han propuesto diversas soluciones. En los DELE, por ejemplo, las pruebas de EIE eran evaluadas de manera independiente por dos calificadores (se repartía un número determinado de pruebas a parejas de calificadores, de forma que cada integrante corrigiera, de manera independiente, los mismos exámenes –original y fotocopia–). Este procedimiento, aunque resulta interesante porque permitía detectar si los dos calificadores de cada bloque de pruebas habían utilizado los criterios de asignación de las calificaciones de manera uniforme (Prieto 2011), no permitía comparar el grado de severidad de todos los calificadores en una misma escala, ya que no estaban relacionados todos los parámetros entre sí.

En el mismo artículo, Linacre y Wright sugieren dos procedimientos diferentes que permiten disminuir el número de calificaciones totales, de forma que el proceso sea viable y se establezca una red entre los parámetros implicados en el proceso: calificadores, candidatos y pruebas, de manera que todos queden relacionados entre sí. En el primero de ellos, aunque se elimina un número considerable de calificaciones, se mantiene la conexión entre candidatos, calificadores y pruebas, al menos dos examinadores califican cada prueba y cada candidato comparte calificador con otro candidato. El ahorro de calificaciones que supone la aplicación de este primer proceso propuesto es del 83%, ya que únicamente es necesario realizar el 17% de las calificaciones que se harían si todos los examinadores calificaran a todos los candidatos (tabla 6). Lógicamente, este ahorro se consigue a costa de disminuir la precisión de las observaciones que se obtienen (Linacre y Wright 2002: figura 2. Cf. también Eckes 2011, Tesio et al. 2015 y Prieto 2015):

Tabla 6

Judge Essay	1 ABC	2 ABC	3 ABC	4 ABC	5 ABC	6 ABC	7 ABC	8 ABC	9 ABC	10 ABC	11 ABC	12 ABC
Person1	553	686										
2		542	445									
3			343	555								
4				545	445							
5					657	448						
6						367	788					
7							773	785				
8								656	784			
9									753	546		
10										667	649	
11											456	334
12	436											676
13	445						368					
14		533						353				
15			743						854			
16				545						447		
17					536						639	
18						473						334
19	747			756								
20		666			557							
21			336			243						
22				666			667					
23					444			388				
24						214			323			
25							454			545		
26								867			756	
27									345			253
28	343									243		
29		444									323	
30			244									555
31	rating performed by any available judges											
32	rating performed by any available judges											

Figura 2 de Linacre y Wright (2002)

El segundo procedimiento que proponen Linacre y Wright para reducir el número de calificaciones totales consiste en que cada una de las tareas de cada prueba de cada candidato sea calificada por un examinador diferente una única vez. De este modo es posible ahorrar el 92% de las calificaciones (únicamente se requieren el 8% de las calificaciones). Sin embargo, aunque se sigue manteniendo la red de calificadores, candidatos y pruebas, la precisión de las calificaciones es más baja que en la propuesta anterior ya que aunque en la calificación de cada prueba de cada candidato participarían tres examinadores (la prueba de EIE del nivel A2 consta de tres tareas) cada tarea sería calificada únicamente por uno. Este procedimiento ha sido implementado con éxito por Lunz, Wright y Linacre en 1990.

Tabla 7

Judge Essay	1 ABC	2 ABC	3 ABC	4 ABC	5 ABC	6 ABC	7 ABC	8 ABC	9 ABC	10 ABC	11 ABC	12 ABC
Person 1	...	...	...	...	...	...	.6.	..7	5..	...	...	...
2	...	...	...	5..	.3.	...	...	...	...	..5	...	...
3	4..	...	...	...	...	...	...	...	...	.5.	..3	...
4	...	.2.	...	...	...	..5	...	...	...	4..	...	...
5	...	...	...	...	..7	...	...	...	...	.4.	5..	...
6	5..	..6	...	...	...	...	...	...	...	...	.6.	...
7	...	...	...	...	.5.	...	..3	...	...	...	...	4..
8	...	.6.	...	...	...	...	...	..6	...	5..	...	...
9	...	...	...	...	.3.	...	6..	..4	...	...	...	...
10	...	...	...	.7.	..5	...	...	...	7..	...	...	...
11	...	...	...	...	...	...	...	...	..6	...	4..	.3.
12	4..	...	...	..6	...	...	...	...	...	...	...	.7.
13	.4.	...	...	..6	...	...	...	...	4..	...	...	...
14	..6	.3.	...	...	...	...	4..	...	...	...	...	...
15	.4.	...	...	...	...	...	...	...	..4	6..	...	...
16	...	6..	.4.	...	...	...	...	...	...	...	..8	...
17	...	...	.2.	6..	..6	...	...	...	...	...	...	...
18	...	...	...	...	...	4..	.4.	...	...	...	...	..4
19	...	7..	...	..6	...	.4.	...	...	...	...	...	...
20	...	...	7..	.5.	...	...	...	...	..6	...	...	...
21	...	...	...	...	...	2..	.6.	...	...	..3	...	...
22	...	...	..7	...	8..	...	...	...	.6.	...	...	...
23	...	..6	7..	...	...	...	...	.8.	...	...	...	...
24	...	...	...	...	...	2..	..3	...	...	...	...	.2.
25	...	...	...	...	...	.4.	..4	...	...	...	4..	...
26	...	...	...	...	...	...	...	...	.7.	..6	...	8..
27	...	...	.3.	...	...	...	...	2..	...	...	..6	...
28	.4.	..3	...	...	...	...	...	...	...	...	...	2..
29	...	...	3..	...	...	...	...	.3.	...	...	...	..3
30	..2	...	...	...	...	...	...	5..	...	...	.3.	...
31	...	...	...	.2.	...	..4	...	2..	...	...	...	...
32	...	...	..5	...	5..	.5.	...	...	...	...	...	...

Figura 3 de Linacre y Wright (2002)

Con el fin de conseguir un equilibrio entre el ahorro y la precisión de las calificaciones, en nuestro análisis hemos seguido el procedimiento indicado en la tabla 6 aunque ligeramente modificado:

Tabla 8

Calificador	1	2	3	4	5	6	7	8	9	10	11	12
Candidato 1	x	x										
2		x	x									
3			x	x								
4				x	x							
5					x	x						
6						x	x					
7							x	x				
8								x	x			
9									x	x		
10										x	x	
11											x	x
12	x		x									
13		x		x								
14			x		x							
15				x		x						
16					x		x					
17						x		x				
18							x		x			
19								x		x		
20									x		x	
21										x		x
22	x			x								
23		x			x							
24			x			x						
25				x			x					
26					x			x				
27						x			x			
28							x			x		
29								x			x	
30									x			x
31	x				x							
32		x				x						
33			x				x					
34				x				x				
35					x				x			
36						x				x		
37							x				x	
38								x				x
39	x					x						
40		x					x					
41			x					x				
42				x					x			
43					x					x		
44						x					x	
45							x					x
46	x						x					
47		x						x				
48			x						x			
49				x						x		
50					x						x	
51						x						x
52	x							x				
53		x							x			
54			x							x		
55				x							x	
56					x							x
57	x								x			
58		x								x		
59			x								x	
60				x								x
61	x									x		
62		x									x	
63			x									x
64	x										x	
65		x										x
66	x											x

A partir del candidato 67 se repite la secuencia.

En la calificación de la prueba participaron 12 calificadores. Cada candidato fue calificado por dos examinadores y se estableció una red en la que quedaron conectados calificadores, pruebas y candidatos.

Los calificadores recibieron una formación general tanto en relación con las exigencias del nivel como en la asignación de las puntuaciones. En concreto, recibieron la siguiente documentación:

- Explicación y ejemplo del examen del nivel A2 (Instituto Cervantes s.f. b).
- Constructo A2.
- Criterios de calificación EIE. Nivel A2 (Instituto Cervantes sf).
- Escalas EIE. Nivel A2 (Instituto Cervantes 2013).
- Instrucciones para los calificadores del nivel A2 (Instituto Cervantes 2013).
- Muestras calificadas del nivel A2 (Instituto Cervantes 2013).
- Manual que detalla el procedimiento que debe seguirse para calificar los exámenes por ordenador.

Las tareas de cada candidato fueron evaluadas independientemente por dos calificadores, que otorgaron sus calificaciones en tres atributos: *Adecuación al género discursivo y coherencia*, *Corrección y alcance* y evaluación *Holística.* Las tareas 1 y 2 corresponden a actividades comunicativas de interacción escritas y la tarea 3 a actividades de expresión escrita.

El atributo *Adecuación al género discursivo y coherencia* se subdivide, a su vez, en *Adecuación al género discursivo* y *Coherencia.* En *Adecuación al género discursivo* se valora cómo el candidato adapta el texto al contexto discursivo, al registro, a los interlocutores, a sus intenciones comunicativas, al canal de producción y recepción, etc. Los rasgos indicadores que se analizan son: alcance (equilibrio entre los rasgos lingüísticos –fórmulas, expresiones, cambios de registro…– y extralingüísticos –comportamientos, relaciones sociales, etc.–), la capacidad para adaptarse al contexto, el nivel de adaptación del discurso a la situación planteada, la cantidad de recursos (expresiones idiomáticas, contactos

sociales…) que se utilizan y la corrección y adecuación a la situación comunicativa. En *Coherencia* se tiene en cuenta el control de los recursos necesarios para establecer relaciones entre el discurso y la situación de comunicación (participantes, circunstancias espacio-temporales, etc.), y para marcar las relaciones entre las unidades de significado dentro del propio texto (conectores, pronombres…). Los rasgos indicadores que se analizan son: la disposición y organización de la información (relaciones lógicas, progresión temática, etc.), el tipo y extensión de los enunciados, el uso de conectores y de sustitutos textuales (referentes, proformas, deícticos…) y la pertinencia de la información.

El atributo *Corrección y alcance* está integrado por *Corrección* y *Alcance*. En *Corrección* se evalúa el conocimiento y capacidad de uso de las categorías gramaticales y de las reglas morfosintácticas. Los rasgos indicadores que se analizan son: la proporción de fórmulas memorizadas o modelos de oraciones frente al uso creativo de reglas gramaticales, la incidencia de los errores en la comunicación, la relación entre errores y el grado de familiaridad con las construcciones empleadas o conocimiento (memorización o ensayo), la presencia o ausencia de errores (cantidad e inconsistencias). En *Alcance* se tiene en cuenta el equilibrio entre los recursos léxicos (palabras, grupos de palabras, fórmulas o expresiones) utilizados y los temas y las situaciones de comunicación. Los rasgos indicadores que se analizan son: la selección léxica (medida en términos de cantidad, precisión y variedad) y la relación entre la capacidad de expresión y el grado de familiaridad, previsibilidad o complejidad de los temas y situaciones.

En la evaluación *Holística* se valora la eficacia del candidato en la realización de las tareas de la prueba. Las dimensiones que se tienen en cuenta son: la eficacia comunicativa (en qué medida resuelve el candidato las tareas de la prueba y cuál es su capacidad para desenvolverse con los temas que trata y en las situaciones en las que participa), la eficacia ejecutiva (en qué medida

demuestra el candidato su capacidad para realizar las tareas que se le piden) y la eficacia lingüística (cómo contribuye el repertorio lingüístico que emplea el candidato y la coherencia informativa de los mensajes que transmite al éxito de la comunicación).

El rendimiento en cada atributo fue puntuado en una escala de 0 a 3. A cada una de las puntuaciones o bandas (0, 1, 2 y 3) corresponde un único descriptor ilustrativo que se compara con la actuación del candidato. En los tres atributos la banda que se califica con 2 puntos es equivalente a la descripción del nivel A2 (Plataforma) del *Marco de referencia*; el valor 3 supone una consecución por encima del nivel; el valor 1 supone la no consecución del nivel a la vista de las respuestas; el valor 0 supone que la prueba está en blanco, que no sigue los puntos de orientación dados, que el candidato escribe información irrelevante que no se ajusta al objetivo planteado o que el texto es ilegible. Este ha sido el procedimiento de puntuación que utilizado para obtener los estadísticos suficientes que usaremos en el próximo capítulo en el análisis con MFRM. El estadístico suficiente para un candidato es la suma de los puntos obtenidos en los criterios y las tareas.

En el Anejo se incluyen las escalas (analítica y holística) utilizadas por los examinadores en el proceso de calificación (Instituto Cervantes 2013).

CAPÍTULO 7

Resultados

Idealmente, si se proporciona la formación adecuada, un equipo de examinadores de la prueba de EIE puede funcionar de manera homogénea y ser capaz de calificar de modo altamente fiable. Las calificaciones de un examinador deberían estar relacionadas con el nivel de competencia de los sujetos, de modo que los candidatos con un alto nivel de dominio recibieran puntuaciones elevadas y los de bajo nivel con calificaciones inferiores. No obstante, resulta complicado conseguir la uniformidad en un equipo amplio de calificadores y, en muchas ocasiones, sus calificaciones pueden presentar una disparidad considerable. El modelo MFRM permite analizar la fiabilidad de un grupo de examinadores y determinar si hay hábitos adquiridos o tendencias idiosincrásicas que afectan al proceso. FACETS permite visualizar el mapa de la variable en una única tabla en la que se presentan los elementos de las diferentes facetas analizadas calibrados en la misma escala de intervalos (*logit*). De este modo es posible comparar e interpretar los resultados de la competencia de los candidatos, la severidad de los calificadores, la dificultad tanto de las tareas como de los atributos, así como la localización de los valores de paso de los valores adyacentes en un mismo marco de referencia.

7.1. Análisis realizado con el modelo RSM

7.1.1. Mapa de la variable. Análisis de las pruebas

En las tablas 9 y 10 se muestran los datos tras analizar las dos pruebas de EIE de los exámenes D031 (tabla 9) y D032 (tabla 10). Es relevante destacar que el equipo de examinadores que participó en el proceso de calificación de las dos

pruebas era el mismo. Las facetas estudiadas son las siguientes: 1 = candidato; 2 = calificador; 3 = tarea, y 4 = atributo.

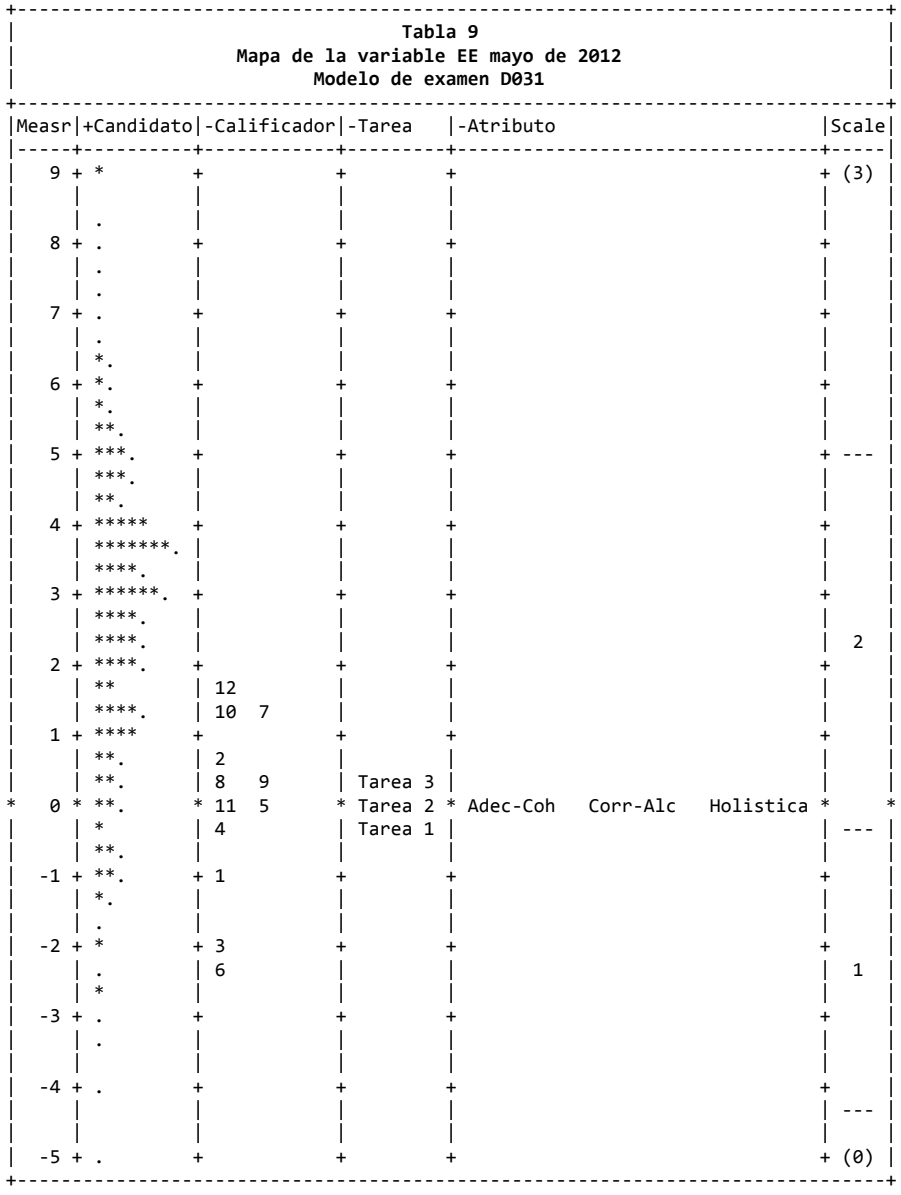
El objetivo de este primer análisis es determinar la localización de las variables dentro del mapa, aunque analizaremos con mayor profundidad la severidad de los calificadores en cada una de las pruebas en relación con la faceta tarea. Para este propósito se aplicará a las cuatro facetas el modelo RSM.

En la primera columna de la tabla, comenzando desde la izquierda, figura la escala *logit* en la que, como es habitual en los modelos Rasch, se suele situar el punto 0 en la dificultad media de las tareas, de los atributos y de la severidad media de los calificadores —únicamente se permite variar libremente la faceta correspondiente a los examinados—. Aunque teóricamente la escala *logit* puede adoptar valores entre 0 $\pm\infty$, en la gran mayoría de los casos se sitúa en el rango $\pm$ 5 (Prieto y Delgado 2003: 95).

En la segunda de las columnas, la de candidatos, se muestra la distribución de esta faceta en la escala de *logit*. Es destacable la elevada variabilidad en la competencia de los candidatos (entre –7,01 y 8,79 en la del examen D031 y entre –4,09 y 10,05 en la del examen D032). También se observa que estos obtienen un rendimiento medio muy superior a la dificultad de las tareas. Es conveniente recordar, tal y como señalábamos en la ecuación 1 del capítulo anterior, que los valores de B_n (capacidad o nivel en la variable latente de la persona) mayores que 0, indican que los candidatos tienen una probabilidad superior a ,50 de superar con éxito las tareas de dificultad media.

En las tablas, los asteriscos (*) representan frecuencias de candidatos. En la tabla 9, cada asterisco representa a tres sujetos y en la tabla 10 a dos. En ambas tablas cada punto (.) representa una frecuencia inferior. Los candidatos con mayor puntuación se sitúan en la parte superior de la tabla, mientras que los de menor puntuación se encuentran en la inferior.

Tabla 9
Mapa de la variable EE mayo de 2012
Modelo de examen D031



*Measr = Medición Rasch (*logit*)*
Scale = Escala en puntuaciones directas (0-3)

Tabla 10
Mapa de la variable EE mayo de 2012
Modelo de examen D032

Measr	+Candidato	-Calificador	-Tarea	-Atributo	Scale
9	**				(3)
8	.				
	.				
	.				
7	***.				
	.				
	*				
6	***				
	**.				
	****.				
5	***				---
	******.				
	****				
4	*****				
	*****				
	****.				
3	*******				
	********				
	*****				2
2	******				
	****.	10			
	***	7			
1	**	2 8			
	*.				
	**.	4 5 9		Corr-Alc	
0	.		Tarea 2 Tarea 3	Adec-Coh	---
	*.		Tarea 1	Holistica	
	*.	11 3			
-1		12			
	*				
		1			
-2					
		6			1
-3	.				
	.				
-4	.				
-5					(0)

*Measr = Medición Rasch (*logit*)*
Scale = Escala en puntuaciones directas (0-3)

En la columna Calificador se visualiza el mapa de los calificadores. Se observa que existe entre ellos una variabilidad en severidad alta (un rango de 2 *logit*, aproximadamente); idealmente, los examinadores apenas deberían mostrar diferencias en esta variable. Esta elevada variabilidad indica que los

calificadores no han utilizado de manera uniforme los atributos de calificación.

En la tabla 9 los calificadores más severos son el 12, el 10 y el 7, mientras que los más benévolos son el 6, el 3 y el 1. En la tabla 10 los examinadores más severos son el 10, el 7, el 2 y el 8 y los más benévolos son el 6 y el 1. Se constata que los calificadores 10 y 7 se encuentran dentro del grupo de los más severos en ambas pruebas, mientras que el 6 y el 1 se hallan entre los más benévolos. El examinador 3, aunque también está en el grupo de los benévolos, es superado en la tabla 10 en benevolencia por el 12. Resulta llamativo lo que sucede con el calificador 12: en la prueba del examen D031 es el más severo mientras que en la del D032 está dentro del grupo de los más benévolos.

En la columna Tarea se muestra el nivel de dificultad de cada una de las tareas que integran las dos pruebas, ordenadas en la escala de *logit* de más difícil (arriba) a más fácil (abajo). La diferencia en los parámetros de dificultad en la prueba del examen D031 entre la tarea más difícil, la 3 (,30 *logit*) y la más fácil, la 1 (–,27 *logit*) es de ,57 *logit*; en la prueba del examen D032 la diferencia es de ,37 *logit* entre la tarea más difícil (en este caso la 2; ,16) *logit* y la más fácil (1; –,21 *logit*).

Situación similar es la que se observa en la columna Atributo, en la que apenas se observan diferencias de dificultad entre los atributos evaluados por los examinadores para calificar las tareas de los candidatos (las diferencias inferiores medio *logit* son irrelevantes desde el punto de vista práctico). Los valores en dificultad oscilan entre los –,14 *logit* de Holística y los ,15 *logit* de Corrección y alcance en el examen D031 (,34 *logit* de diferencia) y entre los –,22 *logit* de Holística y los ,33 *logit* de Corrección y alcance en el examen D032 (,55 *logit* de diferencia). El atributo más difícil en ambas pruebas es Corrección y alcance, seguida de Adecuación al género discursivo y coherencia (en

adelante Adecuación y coherencia); la más fácil, también en las dos pruebas, es la evaluación Holística.

Finalmente, en la columna Escala se indica, mediante pequeñas líneas horizontales, la localización de los valores de paso entre las puntuaciones utilizadas (de 0 a 3) por el conjunto de calificadores en las dos pruebas. En las dos tablas se asume que los pasos no varían entre los examinadores, es decir, que todos los calificadores han utilizado los atributos de evaluación de manera uniforme.

7.1.2. Calificadores

Los mapas de las variables de las dos pruebas anteriormente comentados evidencian que existe una excesiva variabilidad entre las calificaciones otorgadas por los calificadores. En las tablas 11 y 12 se detallan los estadísticos principales de los examinadores.

En la primera columna, comenzando por la izquierda, figura el código de identificación del calificador. En la segunda, los valores en severidad de cada uno de ellos (en *logit*), aspecto que se ha utilizado como referencia para la ordenación de la tabla. En la tercera (SE), el error estándar de la medida (precisión); en la cuarta y quinta, los estadísticos de ajuste (*Infit* y *Outfit*). En la sexta (Rc-rc), la correlación entre las puntuaciones de cada calificador con el resto (concordancia entre calificadores). En la séptima (promedio imparcial o *Fair Average*), se muestra el promedio justo de las evaluaciones de los calificadores (en la escala de 0 a 3). En la octava (promedio observado o *Observed Average*), el promedio observado (también en la escala de 0 a 3). En la novena y última columna, figura el número total de calificaciones que ha asignado el examinador.

Es conveniente recordar que el dato estadístico de la severidad mediante el cual se ha ordenado la tabla se da en la misma escala de intervalos (*logit*) que el resto de facetas analizadas. Además, el procedimiento de calificación utilizado ha permitido establecer una red en la que han quedado conectados calificadores, pruebas y candidatos. Por lo tanto, cada uno de los calificadores en la escala está relacionado con el resto y no únicamente con su "pareja", como sucede en otros tipos de análisis.

Tabla 11
Resumen de los estadísticos de los calificadores
Modelo de examen D031
Ordenado por grado de severidad

Calificador	Severidad	SE	*Infit*	*Outfit*	Rc-rc	Promedio observado	Promedio imparcial	Total calificaciones
12	1,81	,12	,89	,93	,81	1,77	1,76	385
7	1,28	,11	,97	1,00	,57	1,84	1,86	405
10	1,22	,12	,85	,81	,72	1,88	1,87	327
2	,75	,11	,72	,65	,68	1,96	1,94	422
8	,32	,11	1,05	1,04	,66	1,92	1,99	417
9	,29	,11	,90	,91	,85	1,88	2,00	411
5	,01	,11	1,17	1,25	,70	2,02	2,03	383
11	-,09	,11	,91	,87	,63	2,10	2,05	396
4	-,22	,12	1,28	1,22	,57	2,13	2,06	414
1	-1,04	,12	,81	,73	,80	2,21	2,19	401
3	-1,91	,12	1,28	1,25	,73	2,38	2,38	402
6	-2,40	,16	1,21	1,76	,61	2,46	2,51	230

RSR (Rater Separation Reliability): ,99

Tabla 12
Resumen de los estadísticos de los calificadores
Modelo de examen D032
Ordenado por grado de severidad

Calificador	Severidad	SE	*Infit*	*Outfit*	Rc-rc	Promedio observado	Promedio imparcial	Total calificaciones
10	1,70	,15	,66	,62	,74	1,82	1,86	230
7	1,22	,12	1,23	1,27	,58	1,98	1,94	351
8	1,15	,13	1,09	1,11	,65	1,89	1,95	304
2	,88	,15	,75	,69	,69	1,89	1,99	225
4	,29	,13	,79	,75	,55	2,28	2,08	297
5	,28	,13	,90	,84	,73	2,12	2,08	306
9	,28	,13	1,22	1,24	,82	2,10	2,08	297
3	-,52	,14	1,19	1,21	,72	2,16	2,22	277
11	-,55	,14	,98	,95	,60	2,24	2,23	269
12	-,89	,14	1,11	1,04	,63	2,38	2,30	288
1	-1,65	,14	,88	1,21	,69	2,38	2,49	279
6	-2,18	,18	1,01	1,03	,56	2,57	2,62	180

RSR (Rater Separation Reliability): ,99

7.1.2.1. Severidad

En ambas pruebas se observa una elevada variabilidad de los calificadores en severidad, algo más elevada en la prueba del examen D031 (entre –2,40 y 1,81 *logit*) que en la del D032 (entre –2,18 y 1,70 *logit*). Este hecho no es deseable ya que, idealmente, las variaciones en severidad, además de ser mínimas, deberían poder ser atribuidas al error de la medida (Prieto 2011: 236). También el índice RSR es muy elevado en las dos pruebas, lo que revela que las diferencias en severidad observadas entre los calificadores son muy fiables. La precisión de la medición en ambas pruebas es alta, como lo demuestra el hecho de que los errores estándar sean bastante bajos y muy uniformes entre los calificadores (Linacre 2004, 2005).
Con el fin de analizar el comportamiento de los examinadores extremos, agruparemos en cada una de las dos pruebas a los tres calificadores más severos y a los tres más benévolos.

El calificador más severo en la prueba del examen D031 fue el 12 (1,81 *logit*). A continuación se encuentran el 7 (1,28 *logit*) y el 10 (1,22 *logit*). La calificación promedio de estos tres calificadores fue de 1,44 *logit*. Los calificadores más benévolos en la misma prueba fueron el 6 (–2,40 *logit*), el 3 (–1,91 *logit*) y el 1 (–1,04 *logit*). La calificación promedio de estos tres calificadores fue de 1,78 *logit*.

En la prueba del examen D032 el calificador más severo es el número 10, con 1,70 *logit*, seguido por el 7 (1,22 *logit*) y el 8 (1,15 *logit*); la calificación promedio de estos tres calificadores fue de 1,36 *logit*. El calificador más benévolo fue el 6, al igual que en la prueba del examen anterior, con –2,18 *logit* y en penúltimo lugar se encuentra el 1 (–1,65 *logit*). Sorprende la presencia en antepenúltimo lugar —dentro del grupo de los más benévolos— del calificador 12, que en la prueba del examen D031 era el más severo, con –,89 *logit*. La calificación promedio de estos tres calificadores fue de 1,57 *logit*.

Como los calificadores que han intervenido en los procesos de calificación de las dos tareas son los mismos, resulta interesante conocer si existe una correlación, es decir, una asociación o relación entre los valores de severidad de los calificadores de ambas pruebas tanto en términos de direccionalidad como en términos de intensidad o de fuerza.

Para poder utilizar un indicador que permita establecer la covariación entre estas dos variables se utiliza habitualmente el coeficiente de correlación de Pearson. Se trata de una medida de la asociación lineal entre dos variables que oscila entre –1 y +1. El signo del coeficiente indica la dirección de la relación. Si el valor es negativo implica que la correlación es de tipo inverso o indirecto y si es positivo que la correlación es de carácter directo. El valor 0 significa que no existe correlación alguna entre las variables analizadas. El valor absoluto del coeficiente indica la fuerza de la relación entre las dos variables.

El coeficiente de correlación entre las dos variables analizadas es de .67. Habitualmente se considera que si $| r | < ,3$ la asociación es débil, si $| r | \geq ,30$ y $\leq ,70$ la asociación es moderada y si $| r | > 0,70$ significa que la asociación es fuerte lo que significa que la relación existente entre ambas variables es moderada con tendencia a fuerte. En consecuencia, la moderadamente alta correlación entre los valores de severidad de los calificadores en dos exámenes distintos, indica que los examinadores mantienen un criterio o estilo de calificación bastante estable, independientemente de la prueba calificada.

El dato del error estándar se utiliza para establecer los límites probables (intervalos de confianza) entre los que se espera que el valor —en este caso el de la severidad— se encuentre un determinado tanto por ciento de ocasiones si se repitiera el proceso en las mismas condiciones. En el supuesto de una distribución normal, lo esperable es que la medida "verdadera" de severidad

de un calificador se encuentre entre ±2SE en el 95% de las ocasiones (Eckes 2011: 54). Por ejemplo, como la medida de severidad del calificador 12 en la prueba del examen D031 fue de 1,81 *logit* y su SE = ,12, lo esperable es que el potencial error del estadístico no sea superior a ± 2 x 0,12. El intervalo, en ese supuesto, tiene como límite inferior 1,57 *logit* (es decir, 1,81 – 2 x 0,12) y como límite superior 2,05 *logit* (es decir, 1,81 + 2x0,12).

En igualdad de condiciones, como sucede en el análisis realizado, cuanto mayor es el número de calificaciones en las que se basa la estimación, menor será el error estándar y, en consecuencia, más pequeño el intervalo resultante. Por este motivo, como en la prueba del examen D031 la media de pruebas calificadas por examinador fue de 382,75, la media de SE fue de 11,83, mientras que en la prueba del examen D033, como la media de pruebas calificadas por examinador fue de 275,25, la media de SE fue de 14.

Tabla 13
Intervalos de confianza de los calificadores
Modelo de examen D031
Ordenado por grado de severidad

Calificador	Severidad	SE	Intervalos de confianza
12	1,81	,12	1,69 – 1,93
7	1,28	,11	1,17 – 1,39
10	1,22	,12	1,10 – 1,34
2	,75	,11	0,64 – 0,86
8	,32	,11	0,21 – 0,43
9	,29	,11	0,18 – 0,40
5	,01	,11	-0,10 – 0,12
11	-,09	,11	-0,20 – 0,02
4	-,22	,12	-0,34 – -0,10
1	-1,04	,12	-1,16 – -0,92
3	-1,91	,12	-2,03 – -1,79
6	-2,40	,16	-2,56 – -2,24

También el índice RSR es muy elevado en las dos pruebas (,99 en la del examen D031 y ,98 en la del D032), lo que revela que las diferencias en severidad observadas entre los calificadores son muy fiables. Como lo deseable es que no existieran variaciones sustanciales en severidad entre los calificadores, sería preferible que los niveles de RSR fueran bajos.

A continuación se analiza si existe un solapamiento entre los intervalos de estimación del parámetro de severidad de los calificadores. La existencia de un solapamiento entre los intervalos de dos calificadores revela que no hay una diferencia fiable en severidad entre ambos. Existen diversos factores que pueden contribuir a que un calificador tenga tendencia a valorar con mayor severidad o benevolencia; entre otros podemos destacar: experiencia profesional, rasgos de personalidad, actitudes, características demográficas, carga de trabajo y propósito de la evaluación (Eckes 2011: 55).

Tabla 14
Intervalos de confianza de los calificadores
Modelo de examen D032
Ordenado por grado de severidad

Calificador	Severidad	SE	Intervalos de confianza
10	1,70	,15	1,55 – 1,85
7	1,22	,12	1,10 – 1,34
8	1,15	,13	1,02 – 1,28
8	1,15	,13	1,02 – 1,28
2	,88	,15	0,73 – 1,03
4	,29	,13	0,16 – 0,42
5	,28	,13	0,15 – 0,41
9	,28	,13	0,15 – 0,41
3	-,52	,14	-0,66 – -0,38
11	-,55	,14	-0,69 – -0,41
12	-,89	,14	-1,03 – -0,75
1	-1,65	,14	-1,79 – -1,51
6	-2,18	,18	-2,36 – -2,00

En la prueba del examen D031, tal y como se puede observar en la tabla 13, existen ocho grupos de calificadores según sus intervalos de confianza.

En la prueba del examen D032 (tabla 14), también encontramos a los calificadores unidos en ocho grupos según sus intervalos de confianza.

Los estadísticos de ajuste (*Infit* y *Outfit*) indican el grado en que los examinadores utilizan las escalas de calificación de manera consistente y se refieren al grado en que se asocia un determinado calificador con calificaciones inesperadas en relación con los candidatos y los atibutos de calificación. En general, estos estadísticos indican el grado en que las calificaciones observadas coinciden con las esperadas.

Outfit es el término que se utiliza para denominar el ajuste estadístico no ponderado. Es la abreviatura de "outlier-sensitive fit statistic" (ajuste estadístico de valores extremos). (Eckes 2011: 57) y se determina por el promedio de las desviaciones o diferencias cuadráticas estandarizadas entre el desempeño observado y el esperado.

Este estadístico, cuando se utiliza para analizar el comportamiento de los calificadores, es especialmente sensible a calificaciones inesperadas por parte de un calificador que en general califica de manera consistente.

Menos sensible a las calificaciones periféricas inesperadas es el *Infit*. Este estadístico en los calificadores es más sensible que el anterior a las calificaciones inesperadas que no son extremas. Por lo general se considera que las calificaciones no extremas están generalmente asociadas con una mayor precisión de las calificaciones. *Infit* es la abreviatura de "information weighted fit statistic" (información ponderada del ajuste estadístico) (Eckes 2011: 58). Se calcula con el promedio ponderado de las desviaciones o diferencias cuadráticas estandarizadas entre el desempeño observado y el esperado.

El valor esperado de estos dos estadísticos es 1. Cuando los calificadores presentan valores de ajuste superiores a 1 significa que existe mayor variación de lo esperado en sus calificaciones: inadaptación (*misfit* o *underfit*). Si se produce la situación inversa, es decir, que los calificadores tienen valores de ajuste inferiores a 1, es que existe una variación menor que la esperada, lo que indica que sus calificaciones son excesivamente predecibles: sobreajuste (*overfit*). Por lo general, *misfit* es más problemático que *overfit* ya que *misfit* puede cambiar de forma sustantiva las medidas resultantes y poner en peligro la validez del sistema de medición (Wright y Linacre 1994; Myford y Wolfe 2004a).

El establecimiento de los límites inferiores y superiores de los estadísticos de ajuste depende de la naturaleza del propósito de la evaluación. Los exámenes de altas consecuencias, por ejemplo, deben ser más estrictos en el establecimiento de los límites que los de bajas consecuencias.

Linacre (2002b, 2010a, 2003) considera que los límites de los estadísticos de ajuste son 0,5 como límite inferior y 1,5 como límite superior. En otros estudios (Bond y Fox 2007, McNamara 1996, Wright y Linacre 1994) se utiliza un rango más estrecho: entre un límite inferior de 0,70 o 0,75 y uno superior de 1,30. El límite de los valores está relacionado con el tamaño de la muestra. Convencionalmente se considera que con menos de 500 casos indican desajuste los valores superiores a 1,3; con 500–1000 casos el desajuste se produciría a partir de 1,2 y en muestras con más de 1000 casos a partir de 1,1 (Smith, Schumaker y Bush 1998; Prieto Dias 2003: 12). De todos modos, convencionalmente se considera que se produce un desajuste severo con las predicciones del modelo que degrada las medidas cuando *Infit* y/o *Outfit* es superior a 2 (Linacre 2009; citado en Prieto 2011: 234).

Las medias de los estadísticos de *Infit* y de *Outfit* de los calificadores en ambas pruebas, así como de las de sus desviaciones típicas, revelan que los examinadores se ajustan de forma bastante aceptable al modelo. Ninguna de

las medias de los valores de *Infit* y *Outfit* supera ampliamente el valor de 1, que es cuando se manifiesta mayor desajuste del esperado.

Un calificador, el 6, muestra en la prueba del examen D031 un grado bastante elevado de *misfit*, mientras que los calificadores 1 y 2 en la del examen D031 y el 4, el 10 y de nuevo el 2, en la del examen D032, presentan una leve tendencia hacia *overfit*.

Un elevado *misfit* en los calificadores puede indicar que los evaluadores tienen un modo peculiar de calificar o que su comportamiento como calificadores no es demasiado consistente. Por el contrario, un elevado grado de *overfit* puede indicar la existencia de una tendencia central en la calificación o la presencia de efecto de halo (Engelhard 2002; Myford y Wolfe 2004b). También es importante valorar que si las excesivas diferencias entre calificadores son penalizadas por la institución responsable de supervisar el proceso, es posible que indirectamente se favorezcan este tipo de actitudes.

La correlación entre cada calificador con el resto de calificadores (Rc-rc) cuantifica el grado en el que las calificaciones de los examinadores son consistentes con las del resto. Se considera que los valores superiores a 0,30 indican que la evaluación es consistente y que la ordenación que hacen de los candidatos mediante las calificaciones otorgadas es similar a la del resto de examinadores.

El promedio observado y el promedio imparcial están en la misma métrica que la de las escalas de calificación, en las que la puntuación mínima es 0 y la máxima 3. En el presente análisis de los datos de la muestra, los valores de severidad de los calificadores se transforman de nuevo en puntuación bruta, con un límite inferior de 0 (categoría de calificación más baja) y un límite superior de 3 (categoría de calificación más alta). Por lo tanto, los límites del sumatorio de la Ecuación 12 son 0 y 3, respectivamente.

De este modo es posible establecer comparaciones equitativas entre dos calificadores en la misma métrica. Por ejemplo, en la prueba del examen D031 el promedio observado de los calificadores 10 y 9 es el mismo, lo que podría interpretarse como que los dos calificadores aplican las bandas escalas de calificación de forma similar. Sin embargo, si se analiza el promedio imparcial se constata que el calificador 10 es 0,13 puntos más severo que el 9. En la prueba del examen D032 hay tres calificadores con diferente promedio observado, el 4, el 5 y el 9 que, sin embargo, tienen idéntico promedio imparcial, lo que indica que las variaciones que se observan en los promedios observados se deben a que los candidatos que calificaron tenían diverso nivel de competencia mientras que los calificadores aplicaron las bandas de las escalas de calificación con el mismo nivel de severidad.

7.1.2.2. Tendencia central

Resulta interesante analizar la frecuencia con la que el equipo de calificadores utiliza cada una de las bandas de calificación. Para ello se sigue utilizando el modelo en el que se aplica a las cuatro facetas el modelo RSM. Myford y Wolfe (2004a) consideran como evidencias del efecto de tendencia central un bajo índice de fiabilidad (*Rater Separation Reliability*) o valores de *Infit* y *Outfit* extremos. También es una evidencia de este efecto la baja frecuencia de calificaciones en las categorías extremas.

Tabla 15
Uso de las puntuaciones por los calificadores
Modelo de examen D031

Puntuación	Frecuencia absoluta	Frecuencia Relativa	Promedio	*Outfit*	Valor de paso	SE
0	46	1%	-1,65	1,8	--	--
1	739	16%	-,63	1,0	-4,44	,15
2	2769	61%	2,39	1,0	-,42	,05
3	973	21%	5,28	,9	4,86	,05

Tabla 16
Uso de las puntuaciones por los calificadores
Modelo de examen D032

Puntuación	Frecuencia absoluta	Frecuencia Relativa	Promedio	*Outfit*	Valor de paso	SE
0	19	1%	-3,23	1,5	--	--
1	394	12%	-,14	1,0	-4,87	,28
2	1935	60%	2,82	1,0	-,06	,07
3	883	27%	5,44	1,0	4,93	,05

En las tablas 15 y 16 se observa que existe una clara tendencia a abusar de la puntuación 2, mientras que las calificaciones extremas se utilizan con menor frecuencia (entendemos por calificaciones extremas 1 y 3 ya que 0 resulta una anomalía). El porcentaje de utilización de las puntuaciones 1 y 3 en la prueba del examen D031 es del 37%, mientras que en la del examen D032 es del 39%. Aunque aparentemente este hecho provoca un desequilibrio en la dispersión de las calificaciones de los atributos en la escala, es preciso ser prudentes a la hora de extraer conclusiones, ya que un excesivo uso de puntuaciones medias, como sucede en este caso, podría indicar la carencia real de candidatos de muy bajo o muy alto rendimiento (Myford y Wolf 2004b).

Tabla 17
Uso de las puntuaciones por los calificadores
Modelo de examen D031

	Puntuación	Frecuencia absoluta	Frecuencia Relativa	Promedio	*Outfit*	Valor de paso	SE
Calif. 1	0	2	1%	-3,03	,6	--	--
	1	44	11%	-,18	,8	-4,32	,76
	2	209	55%	2,51	,6	-,29	,20
	3	128	23%	5,95	,7	4,60	,15
Calif. 2	0	2	0%	-1,62	1,2	--	--
	1	55	13%	-,86	,8	-5,01	,75
	2	323	77%	2,43	,9	-,95	,18
	3	42		5,57	1,0	5,96	,21
Calif. 3	0	1	0%	,62	1,2	--	--
	1	45	12%	,86	1,1	-3,82	1,02
	2	144	38%	2,35	,7	.36	,20
	3	194	51%	5,14	,9	3,46	,14
Calif. 4	0	12	3%	-,81	2,4	--	--
	1	7	2%	-1,22*	,3	-,49	,38
	2	302	75%	1,53	1,0	-3,14	,30
	3	84	21%	3,11	1,1	3,63	,14
Calif. 5	0	3	1%	-1,81	1,0	--	--
	1	75	20%	-,51	1,1	-4,59	,60
	2	217	57%	2,65	1,3	-1,10	,18
	3	88	23%	4,67	1,2	4,69	,15
Calif. 6	0	2	1%	-2,12	,4	--	--
	1	7	3%	-,50	,7	-2,36	,81
	2	105	46%	2,91	1,9	-1,44	,44
	3	116	50%	5,07	1,2	3,80	,17
Calif. 7	0	--	--	--	--	--	--
	1	92	23%	-3,18	1,2	--	--
	2	286	71$	-,47	1,0	-2,99	,16
	3	27	7%	1,02	1,3	2,99	,22
Calif. 8	0	9	2%	-1,87	2,2	--	--
	1	60	14%	-,61	1,2	-3,68	,38
	2	301	73%	1,67	1,0	-1,17	,17
	3	44	11%	4,09	,9	4,86	,18
Calif. 9	0	7	2%	-2,49	1,2	--	--
	1	128	32%	-1,00	,9	-4,86	,41
	2	174	43%	2,33	,8	,38	,17
	3	93	23%	5,41	,7	4,48	,17
Calif. 10	0	3	1%	-4,04	,8	--	--
	1	71	22%	-,66	,8	-5,07	,67
	2	216	66%	2,13	1,0	-,48	,17
	3	37	11%	5,29	,9	5,55	,22
Calif. 11	0	3	1%	-1,85	,8	--	--
	1	28	7%	-1,06	1,1	-3,75	,60
	2	290	73%	2,47	1,0	-2,05	,22
	3	75	19%	5,69	1,2	5,81	,17
Calif. 12	0	2	1%	-1,99	1,1	--	--
	1	127	34%	-,61	,8	-5,54	,72
	2	202	54%	2,47	,7	,56	,15
	3	45	12%	4,74	,9	4,98	,20

También es posible analizar los efectos de la tendencia central en cada uno de los calificadores de manera individual (Eckes 2011: 63). Es preciso tener

en cuenta que para poder obtener resultados diferenciados para cada uno de los examinadores es preciso utilizar el modelo híbrido número 2.

Tabla 18
Uso de las puntuaciones por los calificadores
Modelo de examen D032

	Puntuación	Frecuencia absoluta	Frecuencia Relativa	Promedio	Outfit	Valor de paso	SE
Calif. 1	0	--	--	--	--	--	--
	1	13	5%	-3,70	,9	--	--
	2	142	53%	1,58	1,4	-2,88	,44
	3	115	43%	3,74	,8	2,88	,16
Calif. 2	0	4	2%	-3,97	,6	--	--
	1	32	14%	-,72	,9	-4,36	,61
	2	173	77%	1,96	,9	-,91	,23
	3	16	7%	3,88	,9	5,27	,28
Calif. 3	0	1	0%	1,98	2,7	--	--
	1	41	15%	,47*	1,1	-4,30	1,03
	2	141	53%	2,34	,7	,17	,21
	3	85	32%	4,87	,9	4,13	,17
Calif. 4	0	--	--	--	--	--	--
	1	1	0%	1,96	,5	--	--
	2	206	72%	2,26	1,1	-4,34	1,01
	3	81	28%	4,27	1,1	4,34	,16
Calif. 5	0	--	--	--	--	--	--
	1	37	12%	-2,17	,9	--	--
	2	187	63%	,23	,8	-2,66	,21
	3	73	25%	3,14	,8	2,66	,18
Calif. 6	0	--	--	--	--	--	--
	1	3	2%	,15	1,1	--	--
	2	67	39%	1,76	1,4	-2,32	,62
	3	101	59%	3,82	,9	2,32	,19
Calif. 7	0	--	--	--	--	--	--
	1	67	19%	-1,48	1,1	--	--
	2	223	64%	-,13	1,3	-2,02	,16
	3	61	17%	1,52	1,0	2,02	,17
Calif. 8	0	1	0%	-1,45	,9	--	--
	1	12	24%	,20	1,4	-5,20	1,01
	2	191	63%	1,99	1,2	,00	,17
	3	40	13%	5,22	,8	5,20	,23
Calif. 9	0	11	4%	-3,17	,9	--	--
	1	58	20%	-,54	,8	-3,63	,38
	2	119	40%	2,19	1,0	,32	,20
	3	109	37%	4,42	,9	3,31	,17
Calif. 10	0	2	1%	-2,57	2,4	--	--
	1	45	20%	-,95	,8	-6,71	,82
	2	164	74%	2,85	,7	-,04	,22
	3	10	5%	4,87	1,0	6,76	,34
Calif. 11	0	--	--	--	--	--	--
	1	12	4%	-1,23	1,0	--	--
	2	181	67%	,94	1,1	-3,01	,33
	3	76	28%	3,30	1,2	3,01	,18
Calif. 12	0	--	--	--	--	--	--
	1	13	5%	-1,15	,9	--	--
	2	141	52%	1,21	,8	-2,34	,33
	3	116	43%	3,18	1,2	2,34	,15

Por medio de este modelo, que combina un modelo PCM aplicado a los calificadores, con un modelo RSM aplicado a los atritutos, es posible obtener información acerca de cómo cada uno de los calificadores utilizan las escalas de calificación en el conjunto de los atributos.

La información obtenida de este modo (tablas 17 y 18) puede resultar más útil que la del análisis grupal, ya que es posible observar el comportamiento de cada uno de los calificadores de forma diferenciada y realizar comparaciones con otros examinadores y con él mismo en la prueba del otro examen.

Los porcentajes de utilización de las puntuaciones extremas 1 y 3 por parte de los examinadores en la pueba del examen D031 (tabla 19) son los siguientes (la media del conjunto de calificadores es del 37%):

Tabla 19
Porcentaje de uso de puntuaciones extremas
Modelo de examen D031

Calificador 1:	34%
Calificador 2:	13%
Calificador 3:	63%
Calificador 4:	23%
Calificador 5:	43%
Calificador 6:	53%
Calificador 7:	25%
Calificador 8:	25%
Calificador 9:	55%
Calificador 10:	33%
Calificador 11:	26%
Calificador 12:	46%

Los tres calificadores que tienen un porcentaje más elevado de utilización de las puntuaciones 1 y 3 son el número 3 con un 63%, el 9 con un 55% y el 6 con un 53%; los cinco que tienen un porcentaje más bajo (infrauso) son el número 2 con un 13%, el 4 con un 23%, el 7 y el 8 con un 25% y el 11 con un 26%.

Los porcentajes de utilización de las mismas puntuaciones extremas 1 y 3 por parte de los examinadores en la pueba del examen D032 (tabla 20) son los siguientes (la media del conjunto de calificadores es del 39%):

Tabla 20
Porcentaje de uso de puntuaciones extremas
Modelo de examen D032

Calificador 1:	48%
Calificador 2:	21%
Calificador 3:	47%
Calificador 4:	28%
Calificador 5:	37%
Calificador 6:	61%
Calificador 7:	36%
Calificador 8:	37%
Calificador 9:	57%
Calificador 10:	25%
Calificador 11:	32%
Calificador 12:	48%

Los dos calificadores que tienen un porcentaje más elevado de utilización de las puntuaciones 1 y 3 son el número 6 con un 61% y el 9 con un 57% ; los tres que tienen un porcentaje más bajo (infrauso) son el número 2 con un 21%, el 10 con un 25% y el 4 con un 28%.

De acuerdo con estos indicadores parece que el calificador 2 presenta claramente efecto de tendencia central (el 4 y el 10 podrían estar cerca de presentarlo).

7.1.2.2.1. Chi-cuadrado

El test *chi-cuadrado* parte de la hipótesis de que todos los candidatos son calificados con el mismo nivel de exigencia, es decir, que todos comparten la misma medida de desempeño, después de considerar el error de la medida (Myford y Wolf 2004b). Un valor de chi-cuadrado no significativo puede

sugerir una tendencia generalizada del equipo de calificadores a la tendencia central.

El valor de chi-cuadrado en la prueba del examen D031 (tabla 21) es de 4636,3 con 258 grados de libertad y en la del examen D031 (tabla 22) es de 2685,3 con 183 grados de libertad. Ambos estadísticos resultan estadísticamente significativos (p<,05), lo que sugiere que ninguno de los dos equipos de calificadores presenta tendencia central.

Tabla 21
Resumen del informe de medición de los candidatos con el modelo RSM
Modelo de examen D031

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	*Infit* MnSq	*Infit* ZStd	*Outfit* MnSq	*Outfit* ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	
36.1	17.7	2.02	2.04	2.52	.58	.99	-.2	1.01	-.2		.28		Mean (Count: 259)
9.2	1.3	.48	.47	2.47	.18	.72	1.4	.84	1.5		.38		S.D. (Population)
9.2	1.3	.48	.48	2.48	.18	.72	1.5	.84	1.5		.38		S.D. (Sample)

With extremes, Model, Populn: RMSE .61 Adj (True) S.D. 2.39 Separation 3.94 Strata 5.59 Reliability .94
With extremes, Model, Fixed (all same) chi-square: 4636.3 d.f.: 258 significance (probability): .00

Tabla 22
Resumen del informe de medición de los candidatos con el modelo RSM
Modelo de examen D032

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	*Infit* MnSq	*Infit* ZStd	*Outfit* MnSq	*Outfit* ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	
38.7	18.0	2.16	2.17	3.37	.58	1.00	.0	1.02	-.1		.27		Mean (Count: 184)
8.0	.4	.44	.43	2.26	.21	.49	1.3	.62	1.3		.35		S.D. (Sample)

With extremes, Model, Populn: RMSE .61 Adj (True) S.D. 2.17 Separation 3.54 Strata 5.05 Reliability .93
With extremes, Model, Fixed (all same) chi-square: 2685.3 d.f.: 183 significance (probability): .00

7.1.2.2.2. Estadísticos de ajuste de las categorías numéricas

El estadístico de ajuste *Outfit*, tal y como se ha indicado anteriormente, es especialmente sensible a los casos atípicos y puede proporcionar información útil para detectar efectos de tendencia central (Myford y Wolf 2004b).

Outfit compara la medida promedio de la aptitud del candidato calculado para cada calificación con la medida de aptitud esperada del candidato. Wright y Linacre (1994) sugieren que un "margen razonable" para considerar que el estadístico se mantiene entre unos márgenes aceptables es de 0,6 a 1,4, aunque advierten que estos límites deben interpretarse como sugerencias y no pueden considerarse como fronteras rígidas (Myford y Wolf 2004a: 495). Cuando *Outfit* es mucho mayor que 1 puede indicar efecto de tendencia central.

Si se analizan del uso de las puntuaciones por los calificadores de las tablas 17 y 18 se pueden observar diferencias significativas en los valores *Outfit* de las calificaciones intermedias de determinados examinadores. En la prueba del examen D031 (tabla 17), el calificador que presenta un *Outfit* más elevado en la puntuación 2 es el 6 (1,9), mientras que el 1 es el que lo tiene más bajo (,6). En la del examen D032 (tabla 18) también es el 6 el que presenta mayor valor de *Outfit* en la calificación 2 (1,4), mientras que el 10 es el que exhibe un índice más bajo (,7).

7.1.2.2.3. Análisis de los pasos de las curvas características de las categorías

Otra de las fuentes de información para detectar efectos de tendencia central a nivel individual es el análisis de las diferencias entre los pasos de las curvas características de las categorías, que puede resultar útil para detectar el efecto

de tendencia central (Myford y Wolf 2004b). Estos umbrales (también denominados valores de paso) representan el punto en el que la probabilidad de que un candidato esté clasificado en una de las dos puntuaciones adyacentes es de un 50%. Cuando un calificador presenta efecto de tendencia central, la media de las diferencias entre los pasos de las categorías es mayor en los calificadores que puntúan con un efecto de tendencia central (Eckes 2011: 65), es decir, los valores de paso se encuentran muy dispersos. Además, habitualmente, los calificadores que presentan efecto de tendencia central no suelen utilizar calificaciones extremas.

Al realizar el análisis con el modelo PCM aplicado a la faceta tarea con un modelo RSM aplicado al resto de tareas tal es posible visualizar en el mapa de la variable los valores de paso del conjunto de calificadores en cada una de las tareas (tablas 23 y 24). Se observa que, en líneas generales, el comportamiento de los calificadores en las tres tareas es bastante homogéneo y no se detectan diferencias importantes en los valores de paso entre las tareas de cada una de las pruebas.

Tabla 23
Mapa de la variable EE mayo de 2012
Modelo de examen D031

```
|Measr|+Candidato|-Calificador|-Tarea   |-Atributo                        | S.1 | S.2 | S.3 |
|-----+----------+------------+---------+---------------------------------+-----+-----+-----|
|   9 + *        +            +         +                                 + (3) + (3) + (3) |
|     |          |            |         |                                 |     |     |     |
|     | .        |            |         |                                 |     |     |     |
|   8 + .        +            +         +                                 +     +     +     |
|     | .        |            |         |                                 |     |     |     |
|     | .        |            |         |                                 |     |     |     |
|   7 + .        +            +         +                                 +     +     +     |
|     | .        |            |         |                                 |     |     |     |
|     | *.       |            |         |                                 |     |     |     |
|   6 + *.       +            +         +                                 +     +     +     |
|     | *.       |            |         |                                 |     |     |     |
|     | **.      |            |         |                                 |     |     |     |
|   5 + ***.     +            +         +                                 +     + --- + --- |
|     | ***.     |            |         |                                 | --- |     |     |
|     | **       |            |         |                                 |     |     |     |
|   4 + *****.   +            +         +                                 +     +     +     |
|     | *******. |            |         |                                 |     |     |     |
|     | ****.    |            |         |                                 |     |     |     |
|   3 + ******.  +            +         +                                 +     +     +     |
|     | ****.    |            |         |                                 |     |     |     |
|     | ***.     |            |         |                                 |  2  |  2  |  2  |
|   2 + ****.    +            +         +                                 +     +     +     |
|     | **.      | 12         |         |                                 |     |     |     |
|     | ****     | 10  7      |         |                                 |     |     |     |
|   1 + ****.    +            +         +                                 +     +     +     |
|     | **.      | 2          |         |                                 |     |     |     |
|     | **.      | 8   9      | Tarea 3 |                                 |     |     |     |
*   0 * **.      * 11  5      * Tarea 2 * Adec-Coh   Corr-Alc   Holistica *     *     *     *
|     | *        | 4          | Tarea 1 |                                 | --- | --- |     |
|     | **.      |            |         |                                 |     |     | --- |
|  -1 + **.      + 1          +         +                                 +     +     +     |
|     | *.       |            |         |                                 |     |     |     |
|     | .        |            |         |                                 |     |     |     |
|  -2 + *        + 3          +         +                                 +     +     +     |
|     | .        | 6          |         |                                 |  1  |     |  1  |
|     | *        |            |         |                                 |     |  1  |     |
|  -3 + .        +            +         +                                 +     +     +     |
|     | .        |            |         |                                 |     |     |     |
|     |          |            |         |                                 |     |     |     |
|  -4 + .        +            +         +                                 +     +     +     |
|     |          |            |         |                                 | --- |     | --- |
|     |          |            |         |                                 |     | --- |     |
|  -5 + .        +            +         +                                 + (0) + (0) + (0) |
|-----+----------+------------+---------+---------------------------------+-----+-----+-----|
|Measr| * = 3    |-Calificador|-Tarea   |-Atributo                        | S.1 | S.2 | S.3 |
```

Measr = Medición Rasch (*logit*)
S.1: Model = ?,?,1,?,R3 ; Tarea: Tarea 1
S.2: Model = ?,?,2,?,R3 ; Tarea: Tarea 2
S.3: Model = ?,?,3,?,R3 ; Tarea: Tarea 3

Tabla 24
Mapa de la variable EE mayo de 2012
Modelo de examen D032

```
|Measr|+Candidato|-Calificador|-Tarea   |-Atributo  | S.1 | S.2 | S.3 |
|-----+----------+------------+---------+-----------+-----+-----+-----|
|   9 + **       +            +         +           + (3) + (3) + (3) |
|     |          |            |         |           |     |     |     |
|     |          |            |         |           |     |     |     |
|   8 + .        +            +         +           +     +     +     |
|     | .        |            |         |           |     |     |     |
|     | .        |            |         |           |     |     |     |
|   7 + ***.     +            +         +           +     +     +     |
|     | .        |            |         |           |     |     |     |
|     | *.       |            |         |           |     |     |     |
|   6 + ***      +            +         +           +     +     +     |
|     | **       |            |         |           |     |     |     |
|     | ****.    |            |         |           |     | --- |     |
|   5 + ***.     +            +         +           + --- +     +     |
|     | ******   |            |         |           |     |     | --- |
|     | ****.    |            |         |           |     |     |     |
|   4 + ****.    +            +         +           +     +     +     |
|     | *****    |            |         |           |     |     |     |
|     | *****    |            |         |           |     |     |     |
|   3 + *******  +            +         +           +     +     +     |
|     | *******. |            |         |           |  2  |  2  |     |
|     | *****.   |            |         |           |     |     |  2  |
|   2 + ******   +            +         +           +     +     +     |
|     | ****     | 10         |         |           |     |     |     |
|     | ***      | 7          |         |           |     |     |     |
|   1 + **       + 2   8      +         +           +     +     +     |
|     | **       |            |         |           |     |     |     |
|     | **       | 4   5   9  | Tarea 3 | Corr-Alc  |     |     |     |
*   0 * .        *            * Tarea 2 * Adec-Coh  * --- * --- *     *
|     | *.       |            | Tarea 1 | Holistica |     |     | --- |
|     | *.       | 11  3      |         |           |     |     |     |
|  -1 +          + 12         +         +           +     +     +     |
|     | *        |            |         |           |     |     |     |
|     |          | 1          |         |           |     |     |     |
|  -2 +          +            +         +           +     +     +     |
|     |          | 6          |         |           |  1  |     |  1  |
|     |          |            |         |           |     |  1  |     |
|  -3 + .        +            +         +           +     +     +     |
|     |          |            |         |           |     |     |     |
|     | .        |            |         |           |     |     |     |
|  -4 + .        +            +         +           +     +     +     |
|     |          |            |         |           |     |     | --- |
|     |          |            |         |           |     |     |     |
|  -5 +          +            +         +           + (0) + (0) + (0) |
|-----+----------+------------+---------+-----------+-----+-----+-----|
 |Measr|+Candidato|-Calificador|-Tarea   |-Atributo  | S.1 | S.2 | S.3 |
```

Measr = Medición Rasch (*Logit*)
S.1: Model = ?,?,1,?,R3 ; Tarea: Tarea 1
S.2: Model = ?,?,2,?,R3 ; Tarea: Tarea 2
S.3: Model = ?,?,3,?,R3 ; Tarea: Tarea 3

Tabla 25
Uso de las puntuaciones por los calificadores
Modelo de examen D031

Puntuación	Frecuencia absoluta			Frecuencia Relativa			Promedio			*Outfit*			Valor de paso			SE		
	Tarea			Tarea			Tarea			Tarea			Tarea			Tarea		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
0	14	12	20	1%	1%	1%	-1,74	-,45	-2,23	1,4	2,9	1,3	--	--	--	--	--	--
1	243	241	255	16%	16%	17%	-,46	-,60	-,84	1,0	1,0	1,0	-4,36	-4,60	-4,36	,30	,31	,25
2	859	943	967	57%	62%	64%	2,48	2,45	2,23	,9	,9	1,2	-,24	-,41	-,62	,09	,09	,09
3	395	317	261	26%	21%	17%	5,22	5,46	5,16	1,0	,9	,9	4,60	5,01	4,98	,08	,08	,09

Tabla 26
Uso de las puntuaciones por los calificadores
Modelo de examen D032

Puntuación	Frecuencia absoluta			Frecuencia Relativa			Promedio			*Outfit*			Valor de paso			SE		
	Tarea			Tarea			Tarea			Tarea			Tarea			Tarea		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
0	5	5	9	0%	0%	1%	-3,52	-4,18	-2,66	1,1	,9	2,1	--	--	--	--	--	--
1	123	142	129	11%	13%	12%	,30	,28	-,10	1,1	1,1	,9	-5,03	-5,28	-4,45	,53	,52	,43
2	623	666	646	58%	62%	60%	3,01	2,90	2,64	1,2	,8	1,1	,09	,08	-,28	,12	,11	,12
3	327	267	289	30%	25%	27%	5,56	5,68	5,16	1,0	,9	1,0	4,94	5,20	4,73	,09	,09	,09

Al utilizar este modelo híbrido es posible ampliar la información analizada en las tablas 17 y 18. En las tablas 25 y 26 se observa que la tarea en la que con menor frecuencia relativa se utiliza la puntuación 2 es la primera (57% en la prueba del examen D031 y 58% en la del D032) y que esta calificación se utiliza con más habitualmente en la tarea 3 que en la 2 en la prueba del examen D031, mientras que en la del D032 se utiliza más en la 2 que en la 3.

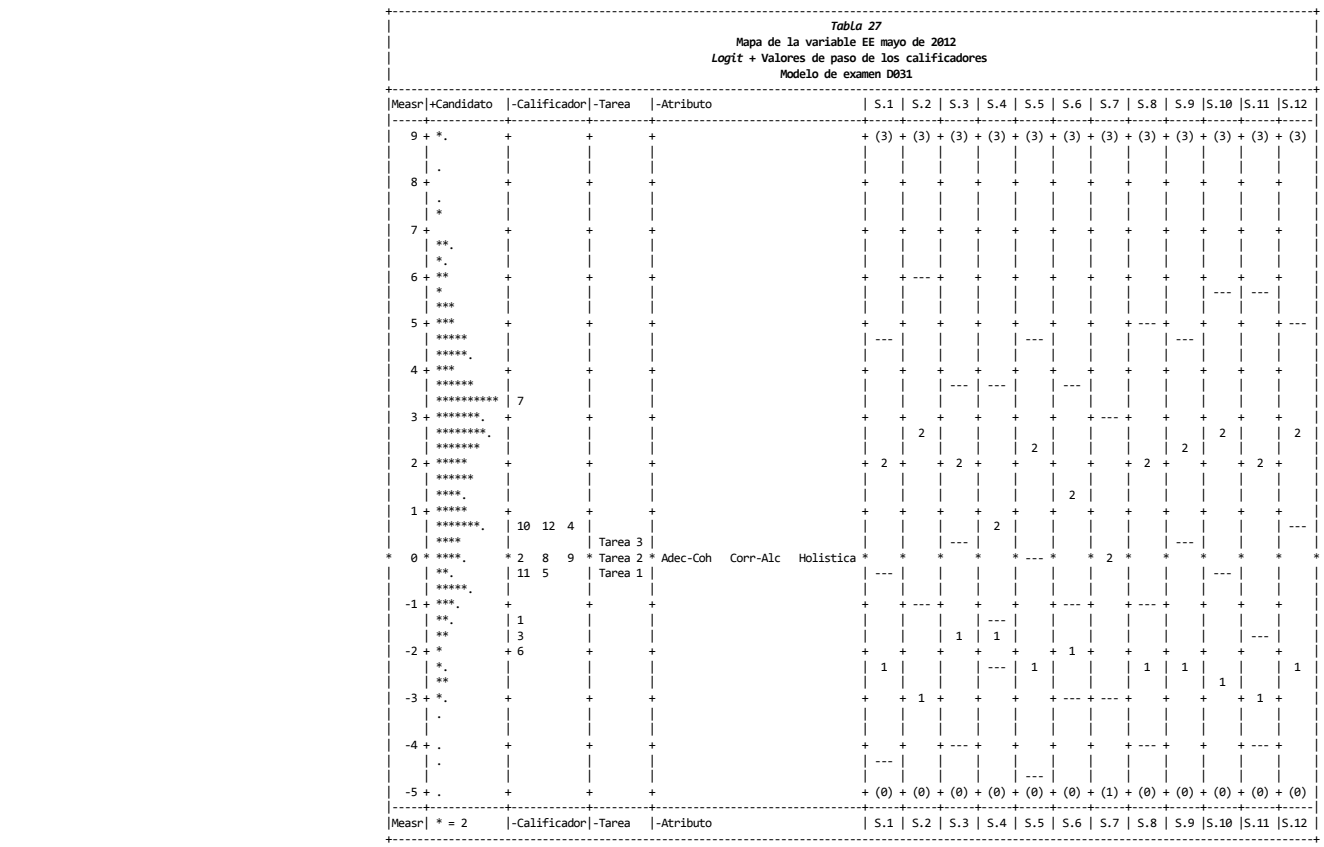


Tabla 27
Mapa de la variable EE mayo de 2012
Logit + Valores de paso de los calificadores
Modelo de examen D031

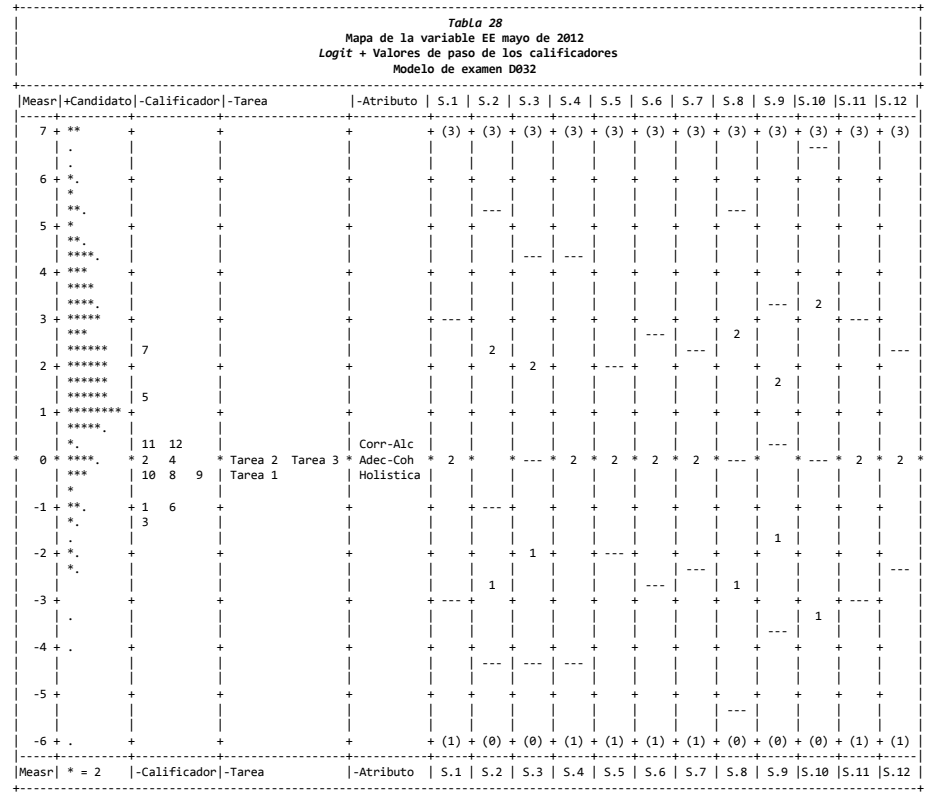
Measr	+Candidato	-Calificador	-Tarea	-Atributo	S.1	S.2	S.3	S.4	S.5	S.6	S.7	S.8	S.9	S.10	S.11	S.12
9	*				(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)
	.															
8																
	.															
	*															
7																
	**.															
	*.															
6	**					---								---		
	*															
	***															
5	***												---			---
	*****				---				---							
	*****.												---			
4	***															
	******						---	---		---						
	**********	7														
3	*******.											---				
	********.					2								2		2
	*******								2				2			
2	*****				2		2					2			2	
	******															
	****.									2						
1	*****															
	*******.	10 12 4						2								---
	****		Tarea 3				---						---			
0	****.	2 8 9	Tarea 2	Adec-Coh Corr-Alc Holistica					---		2					
	**.	11 5	Tarea 1		---									---		
	*****.															
-1	***.					---				---			---			
	**.	1						---								
	**	3					1	1							---	
-2	*	6							1							
	*.				1			---	1			1	1			1
	**													1		
-3	*.					1				---	---				1	
	.															
-4	.						---					---			---	
	.				---											

-5	.				(0)	(0)	(0)	(0)	(0)	(0)	(1)	(0)	(0)	(0)	(0)	(0)
Measr	* = 2	-Calificador	-Tarea	-Atributo	S.1	S.2	S.3	S.4	S.5	S.6	S.7	S.8	S.9	S.10	S.11	S.12

*Measr = Medición Rasch (*logit*)*
S.1, S.2, S.3… = Calificador 1, Calificador 2, Calificador 3…

Sin embargo, el análisis realizado no permite analizar la posible existencia de efectos de tendencia central de manera individualizada en cada uno de los calificadores. Para ello el preciso utilizar el mismo análisis que ya se ha utilizado al comienzo de este apartado y que se resumía en las tablas 11 y 12, en el que para analizar los efectos de la tendencia central se combinaba un modelo PCM aplicado a la faceta calificador con un modelo RSM aplicado al resto de tareas. De este modo, es posible visualizar en el mapa de la variable los valores de paso de cada uno de los calificadores en el conjunto de tareas.

Tabla 28
Mapa de la variable EE mayo de 2012
***Logit* + Valores de paso de los calificadores**
Modelo de examen D032



*Measr = Medición Rasch (*logit*)*
S.1, S.2, S.3… = Calificador 1, Calificador 2, Calificador 3…

En las columnas de la derecha de los mapas de las variables de las tablas 27 y 28 (S.1, S.2, S3…) se visualiza la localización y la amplitud de los valores de paso entre las puntuaciones utilizadas (de 0 a 3) para cada uno de los calificadores. En ambas tablas se asume que los calificadores usan de la misma forma las categorías en todos los atributos.

En general, se considera que cuanto mayor es la media de las diferencias entre los pasos, más posibilidades hay de que el calificador muestre signos de tendencia central. La distancia media entre los valores de paso del calificador 2 en la prueba del examen D031 es de 5,48 *logit*, mientras que la del calificador 7 es de 2,99 *logit* (tabla 29). Esta mayor dispersión que presenta el calificador 2 de las categorías puede evidenciar la presencia de un efecto de tendencia central en este examinador.

En la prueba de este examen dos de los valores de paso entre puntuaciones del calificador 4 son incongruentes. El paso entre las categorías adyacentes 0-1 está ubicado en el mapa de la variable 2.65 *logit* por encima del paso entre las categorías 1-2 (tabla 17).

Tabla 29
Distancia entre valores de paso
Modelo de examen D031

Calificador	Distancia entre valores de paso (en *logit*)	Distancia media entre valores de paso (en *logit*)
7	5,98	2,99
6	6,16	3,08
4	6,77	3,39
3	7,28	3,64
8	8,54	4,27
1	8,92	4,46
5	9,28	4,64
9	9,34	4,67
11	9,56	4,78
12	10,52	5,26
10	10,62	5,31
2	10,97	5,48

En la prueba del examen D032, la media entre los valores de paso del calificador 10 es de 6,74 *logit*, mientras que la del calificador 7 es de 2,02 *logit* (tabla 30). Esta diferencia, cercana a 5 *logit*, indica que el calificador 10 incluye una gama más amplia de niveles de competencia de candidatos en las dos puntuaciones medias de la escala de calificación que el 7. Al igual que indicábamos arriba, la mayor dispersión que presenta el calificador 10 de las categorías puede evidenciar también la presencia de un efecto de tendencia central en este examinador.

Tabla 30
Distancia entre valores de paso
Modelo de examen D032

Calificador	Distancia entre valores de paso (en *logit*)	Distancia media entre valores de paso (en *logit*)
7	4,04	2,02
6	4,64	2,32
12	4,68	2,34
5	5,32	2,66
1	5,76	2,88
11	6,02	3,01
9	6,94	3,47
3	8,43	4,22
4	8,68	4,34
2	9,63	4,83
8	10,40	5,20
10	13,47	6,74

El valor de paso de la puntuación 2 tiene especial relevancia en la prueba objeto de estudio ya que se trata de un valor de paso crítico. Cuando un examinador asigna la calificación 1 a un candidato, significa que considera que este no ha alcanzado los objetivos mínimos descritos para este nivel; en el momento en que el calificador comienza a utilizar la calificación 2 (valor de paso entre las puntuaciones adyacentes 1–2) es cuando, según su criterio, el candidato alcanza el nivel A2, que es el que describe la banda 2 de cada uno de los atributos de calificación.

En la tabla 31 se presentan ordenados los examinadores según la localización de este valor umbral en el mapa de la variable en cada una de las dos pruebas.

Tabla 31
Uso de la calificación 2 por los calificadores

Modelo de examen D031		Modelo de examen D032	
Calificador	**P-2**	**Calificador**	**P-2**
12	,56	9	,32
9	,38	3	,17
3	,36	8	,00
5	,10	10	-,04
1	-,29	2	-,91
10	-,48	7	-2,02
2	-,95	6	-2,32
8	-1,17	12	-2,34
6	-1,44	5	-2,66
11	-2,05	1	-2,88
7	-2,99	11	-3,01
4	-3,14	4	-4,34

P-2: Paso entre las calificaciones 1–2

Aunque se observan diferencias en la ordenación de los calificadores en cada una de las pruebas, sí es posible observar comportamientos similares entre los calificadores. El valor de paso de la categoría 2 del calificador 4 en las dos pruebas se sitúa en la posición más baja en el mapa de la variable. Esto significa que un candidato que es calificado por este examinador, para obtener la calificación 2 no necesitaría demostrar que tiene tanto nivel de competencia como si fuera calificado por cualquier otro examinador. En el calificador 12, por el contrario, el comienzo de la asignación de la categoría 2 en la prueba del examen D031 se localiza en el mapa de la variable mucho más arriba que en el caso del calificador anteriormente analizado, el 4, en el mapa de la variable (,56 *logit*).

Resulta sorprendente el comportamiento de algunos calificadores a la hora de comenzar a asignar la calificación de 2. El valor de paso del calificador 12 en la prueba del examen D031, tal y como habíamos indicado arriba, se

localizaba en la posición más elevada del mapa de la variable, mientras que en la del examen D032 se ubica en una posición condiderablemente inferior.

Además de la localización en el mapa del valor logit a partir del cual el calificador comienza a asignar la calificación 2, es relevante la amplitud en *logit* de la misma. Se considera que cuanto mayor es la distancia entre los pasos que limitan la categoría 2, mayor será la presencia del efecto de tendencia central. Los mayores valores corresponden a los calificadores 11 y 2. También en este caso se considera que cuanto mayor es la distancia los valores de paso intermedios, es posible afirmar que el calificador presenta signos de tendencia central.

Tabla 32
Distancia entre valores de paso

Modelo de examen D031		**Modelo de examen D032**	
Calificador	**Distancia entre valores de paso puntuación 2 (en *logit*)**	**Calificador**	**Distancia entre valores de paso puntuación 2 (en *logit*)**
3	3,82	9	2,99
5	4,79	3	3,96
9	4,86	7	4,04
1	4,89	6	4,64
6	5,24	12	4,68
12	5,54	8	5,20
7	5,98	5	5,32
8	6,03	1	5,76
10	6,03	11	6,02
4	6,77	2	6,18
2	6,91	10	6,80
11	7,86	4	8,68

7.1.2.3. Efecto de halo

7.1.2.3.1. Medidas de los atributos evaluados

Las medias de los atributos evaluados por los calificadores apenas presenta diferencias (tablas 33 y 34). En la prueba del examen D031 la media de Correción y alcance es de 2,01, la de Adecuación y coherencia de 2,03 y la de calificación Holística 2,05. En la del D032 la media de Correción y alcance es de 2,08, la de Adecuación y coherencia de 2,16 y la de calificación Holística 2,18. Es relativa igualdad de medias puede indicar la presencia de efecto de halo en la mayoría de los calificadores.

Tabla 33
Resumen del informe de medición de los candidatos con el modelo RSM
Modelo de examen D031

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	*Infit* MnSq	*Infit* ZStd	*Outfit* MnSq	*Outfit* ZStd	Estim. Discrm	Correlation PtMea	PtExp	N Atributo
3115	1526	2.03	2.04	-.01	.06	1.15	3.6	1.18	3.2	.85	.75	.75	1 Adec-Coh
3071	1524	2.01	2.02	.15	.06	.95	-1.1	1.01	.1	1.04	.72	.75	2 Corr-Alc
3172	1543	2.05	2.05	-.14	.06	.88	-3.1	.85	-2.9	1.12	.78	.75	3 Holistica
3119.3	1531.0	2.03	2.03	.00	.06	.99	-.2	1.01	.1		.75		Mean (Count: 3)
41.3	8.5	.02	.02	.12	.00	.11	2.9	.13	2.5		.03		S.D. (Population)
50.6	10.4	.02	.02	.15	.00	.14	3.5	.16	3.1		.03		S.D. (Sample)

Model, Populn: RMSE .06 Adj (True) S.D. .10 Separation 1.73 Strata 2.64 Reliability .75
Model, Fixed (all same) chi-square: 12.0 d.f.: 2 significance (probability): .00

Tabla 34
Resumen del informe de medición de los candidatos con el modelo RSM
Modelo de examen D032

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	*Infit* MnSq	*Infit* ZStd	*Outfit* MnSq	*Outfit* ZStd	Estim. Discrm	Correlation PtMea	PtExp	N Atributo
2400	1102	2.16	2.14	-.11	.07	1.11	2.5	1.08	1.3	.89	.71	.72	1 Adec-Coh
2307	1100	2.08	2.07	.33	.07	.95	-1.2	.99	-.1	1.05	.72	.72	2 Corr-Alc
2422	1101	2.18	2.16	-.22	.07	.94	-1.4	.97	-.5	1.07	.73	.72	3 Holistica
2376.3	1101.0	2.14	2.12	.00	.07	1.00	-.1	1.01	.2		.72		Mean (Count: 3)
49.8	.8	.05	.04	.24	.00	.08	1.8	.05	.8		.01		S.D. (Population)
61.0	1.0	.06	.05	.29	.00	.10	2.2	.06	1.0		.01		S.D. (Sample)

Model, Populn: RMSE .07 Adj (True) S.D. .23 Separation 3.26 Strata 4.67 Reliability .91
Model, Fixed (all same) chi-square: 35.0 d.f.: 2 significance (probability): .00

7.1.2.3.2. Fiabilidad del índice de separación entre atributos

Este indicador proporciona información acerca de si los atributos se diferencian entre sí y muestra el grado en que los calificadores han sido capaces de distinguirlos. Una baja fiabilidad del índice podría sugerir efecto de halo en las calificaciones (Myford y Wolf 2004b).

La fiabilidad de diferenciación entre atributos es de ,75 en la prueba del examen D031 (tabla 33) y de ,91 en la del D032 (tabla 34). El índice de la prueba del examen D031 quizá podría indicar un cierto efecto de halo que podría implicar que los calificadores no han podido distinguir claramente entre atributos.

Para cada calificador incluido en el análisis, FACETS proporciona medidas de la consistencia entre las calificaciones observadas y las esperadas en los diversos atributos evaluados. Cuando un calificador exhibe efecto de halo asignará a los candidatos calificaciones prácticamente idénticas en atributos diferentes y sus calificaciones mostrarán poca desviación en relación con las esperadas, de modo que sus valores *Infit* y *Outfit* serán significativamente inferiores a 1. Este hecho indicaría que estos calificadores no han sido capaces de diferenciar con fiabilidad entre atributos conceptualmente diferentes (inadaptación).

Tabla 35
Informe de calificaciones de los candidatos con el modelo RSM
Modelo de examen D031

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	*Infit* MnSq	*Infit* ZStd	*Outfit* MnSq	*Outfit* ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Num	Calificador
900	401	2.21	2.19	-1.04	.12	.81	-2.7	.73	-2.3	1.20	.80	.74	54.2	55.6	9	1
827	422	1.96	1.94	.75	.11	.72	-3.7	.65	-4.1	1.23	.68	.66	59.8	61.5	146	2
969	402	2.38	2.38	-1.91	.12	1.28	3.5	1.25	2.0	.66	.73	.72	39.8	45.1	532	3
890	414	2.13	2.06	-.22	.12	1.28	3.2	1.22	2.2	.81	.57	.62	66.9	64.7	800	4
773	383	2.02	2.03	.01	.11	1.17	2.3	1.25	2.6	.80	.70	.72	62.5	60.2	949	5
745	405	1.84	1.86	1.28	.11	.97	-.3	1.00	.0	1.01	.57	.65	57.6	57.0	957	7
565	230	2.46	2.51	-2.40	.16	1.21	2.2	1.76	2.8	.66	.61	.66	40.4	42.4	1031	6
794	417	1.92	1.99	.32	.11	1.05	.6	1.04	.5	.97	.66	.68	62.8	59.0	1085	8
782	411	1.88	2.00	.29	.11	.90	-1.4	.91	-1.0	1.08	.85	.82	56.7	60.6	1120	9
614	327	1.88	1.87	1.22	.12	.85	-1.9	.81	-2.1	1.17	.72	.70	56.6	54.0	1121	10
833	396	2.10	2.05	-.09	.11	.91	-1.2	.87	-1.5	1.11	.63	.70	55.9	59.6	1142	11
666	385	1.77	1.76	1.81	.12	.89	-1.2	.93	-.6	1.08	.81	.78	51.9	52.6	1193	12
779.8	382.8	2.05	2.05	.00	.12	1.00	-.1	1.04	-.1		.69				Mean (Count: 12)	
113.8	51.9	.21	.21	1.21	.01	.18	2.3	.29	2.1		.09				S.D. (Population)	
118.9	54.2	.22	.21	1.27	.01	.19	2.4	.30	2.2		.09				S.D. (Sample)	

Model, Populn: RMSE .12 Adj (True) S.D. 1.21 Separation 10.01 Strata 13.67 Reliability (not inter-rater) .99
Model, Fixed (all same) chi-square: 1059.7 d.f.: 11 significance (probability): .00
Inter-Rater agreement opportunities: 2244 Exact agreements: 1258 = 56.1% Expected: 1271.3 = 56.7%

Según los estudios ya citados de Bond y Fox 2007, McNamara 1996 y Wright y Linacre 1994, el límite inferior de los estadísticos *Infit* y *Outfit* para un examen de altas características se situaría entre 0,70 o 0,75. En la prueba del examen D031 (tabla 35) presenta medidas en *Infit* y/o *Outfit* inferiores a las indicadas el calificador 2 (el valor del *Outfit* del calificador 1 estaría muy cerca del límite); en la del examen D032 (tabla 36) son los calificadores 2 y 10 los que tienen valores inferiores a los indicados. En ninguna de las pruebas hay calificadores con valores *Infit* /*Outfit* superiores a 1,30 (cf. 7.1.2.2.2.).

Tabla 36
Informe de calificaciones de los candidatos con el modelo RSM
Modelo de examen D032

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	*Infit* MnSq	*Infit* ZStd	*Outfit* MnSq	*Outfit* ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Num	Calificador
669	279	2.38	2.49	-1.65	.14	.88	-1.5	1.21	1.6	1.10	.69	.68	51.1	52.0	9	1
426	225	1.89	1.99	.88	.15	.75	-2.5	.69	-2.8	1.24	.69	.66	60.0	57.4	146	2
605	277	2.16	2.22	-.52	.14	1.19	2.1	1.21	1.9	.78	.72	.74	56.6	57.9	532	3
683	297	2.28	2.08	.29	.13	.79	-2.9	.75	-2.8	1.25	.55	.58	57.6	58.6	800	4
657	306	2.12	2.08	.28	.13	.90	-1.3	.84	-1.6	1.14	.73	.70	62.8	60.6	949	5
696	351	1.98	1.94	1.22	.12	1.23	2.7	1.27	2.8	.79	.58	.57	52.4	56.5	957	7
467	180	2.57	2.62	-2.18	.18	1.01	.1	1.03	.1	.97	.56	.56	48.5	47.7	1031	6
574	304	1.89	1.95	1.15	.13	1.09	1.0	1.11	1.1	.90	.65	.71	53.9	57.0	1085	8
623	297	2.10	2.08	.28	.13	1.22	2.5	1.24	2.2	.76	.82	.80	54.9	60.5	1120	9
430	230	1.82	1.86	1.70	.15	.66	-3.9	.62	-4.0	1.33	.74	.71	60.0	52.6	1121	10
602	269	2.24	2.23	-.55	.14	.98	-.1	.95	-.4	1.03	.60	.66	59.9	62.5	1142	11
697	288	2.38	2.30	-.89	.14	1.11	1.4	1.04	.3	.86	.63	.65	52.2	59.6	1193	12
594.1	275.3	2.15	2.15	.00	.14	.98	-.2	1.00	-.1		.66				Mean (Count: 12)	
96.3	43.1	.22	.22	1.14	.02	.18	2.2	.22	2.2		.08				S.D. (Population)	
100.5	45.0	.23	.23	1.19	.02	.19	2.3	.23	2.3		.08				S.D. (Sample)	

Model, Populn: RMSE .14 Adj (True) S.D. 1.13 Separation 7.96 Strata 10.95 Reliability (not inter-rater) .98
Model, Fixed (all same) chi-square: 699.0 d.f.: 11 significance (probability): .00
Inter-Rater agreement opportunities: 1612 Exact agreements: 901 = 55.9% Expected: 924.1 = 57.3%

7.1.3. Tareas

En las tablas 37 y 38 figuran los estadísticos de las medidas de dificultad de las tres tareas de las dos pruebas.

El comportamiento de las tareas es similar en ambos exámenes: la tarea 1 es la más fácil de las tres (–,27 *logit* en el examen D031 y –,21 *logit* en el examen D032); la segunda en nivel de dificultad es la 2 (,03 *logit* en el examen D031 y ,05 *logit* en el examen D032) y la más difícil es la 3 (,30 *logit* en el examen D031 y ,16 *logit* en el examen D032). El error típico de la medida (SE) es muy bajo en ambas pruebas (,06 en la del examen D031 y ,07 en la del D032). Este hecho indica que la precisión de las estimaciones de la dificultad es muy alta. El ESR, el estadístico de fiabilidad global de las estimaciones de la dificultad de los ítems, es elevado en ambos casos (especialmente en el examen D031). Esto sucede cuando las tareas del examen tienen diferente dificultad porque intentan garantizar un muestreo adecuado de los distintos niveles del constructo que se evalúan. Los estadísticos de ajuste *Infit* y *Outfit* presentan

unos valores muy próximos a 1 (valor esperado), lo que evidencia que ninguna de las tareas se desajusta severamente con las predicciones del modelo.

Tabla 37
Valores y estadísticos de las tareas
Modelo de examen D031

Tarea	Promedio	Dificultad	SE	*Infit*	*Outfit*	RiX
1	2.08	–,27	,06	1,08	1,07	,75
2	2,03	,03	,06	,97	,97	,75
3	1,98	,30	,06	,93	,99	,75
Media	*2,03*	*,00*	*,06*	*,99*	*1,01*	*,75*
DT	*,05*	*,29*	*,00*	*,08*	*,05*	*,00*

ESR (Exercise Separation Reliability) = ,94

Tabla 38
Valores y estadísticos de las tareas
Modelo de examen D032

Tarea	Promedio	Dificultad	SE	*Infit*	*Outfit*	RiX
1	2,18	–,21	,07	1,03	1,09	,72
2	2,13	,05	,07	1,04	1,05	,72
3	2,11	,16	,07	,93	,90	,73
Media	*2,14*	*,00*	*,07*	*1,00*	*1,01*	*,72*
DT	*,04*	*,19*	*,00*	*,06*	*,10*	*,01*

ESR (Exercise Separation Reliability) = ,84

7.1.4. Candidatos

En la tabla 39 se detallan los principales estadísticos de las puntuaciones de los candidatos que realizaron las pruebas de ambos exámenes. El modelo utilizado en este análisis es el modelo RSM. Se expresa el rendimiento de los candidatos como puntuación directa obtenida en la escala 0–3 (X) y en la escala *logit* (MFRM).

En la variable X el rendimiento promedio del conjunto de los candidatos es elevado, superior a 2 (de un máximo de 3) en las dos pruebas, aunque algo menor en la del examen D031 (2,02) que en la del D032 (2,16). Las

desviaciones típicas de las puntuaciones también son altas (,48 en la prueba del examen D031 y ,44 en la del D032), lo que indica una gran variabilidad en el rendimiento de los examinados. De igual modo, en la escala *logit* se aprecia un rendimiento medio bastante superior a la dificultad media de las tareas: 2,52 en la prueba del examen D031 y 3,37 en la del D032.

Tabla 39
Resumen de los estadísticos de los candidatos con el modelo RSM

	Modelo de examen D031		Modelo de examen D032	
Variable	**X**	***Logit***	**X**	***Logit***
Media	2,02	2,52	2,16	3,37
Desviación típica	,48	2,47	,44	2,25
Máximo	3,00	8,79	3,00	10,05
Mínimo	0,00	-7,01	,92	-4,09
SE promedio	——	,58	——	,58
Fiabilidad (PSR)	——	,94	——	,93
Media de *Infit*	——	,99	——	1,00
DT *Infit*	——	,72	——	,49
Media de *Outfit*	——	1,01	——	1,02
DT *Outfit*	——	,84	——	,62

X = puntuación directa en la escala 0-3

Las desviaciones típicas de los examinados en la escala *logit* son elevadas en ambos exámenes: 2,47 y 2,25 respectivamente. La gran variabilidad se manifiesta asimismo en el rango (variación entre la mayor y la menor puntuación): entre –7,01 *logit* y 8,79 *logit* en la prueba del examen D031 y entre –4,09 *logit* y 10,05 *logit* en la del D032.

Las puntuaciones otorgadas a los candidatos tienen una fiabilidad muy alta: PSR = ,94 *logit* en la prueba del examen D031 y ,93 *logit* en la del examen D032. El estadístico PSR es comparable al coeficiente alfa utilizado en la Teoría Clásica de los Tests: oscila entre 0 y 1 y puede interpretarse como el grado en que las puntuaciones en la prueba de EIE permiten diferenciar con fiabilidad entre los niveles de competencia de los candidatos (Prieto 2011: 334). La media de los errores estándares de la medida (SE) es la misma en las dos pruebas ,58 *logit*. Se utiliza este estadístico para estimar el intervalo de puntuación entre el que se encuentra la puntuación verdadera de los

candidatos. Es decir, hay un 95% de confianza en que la competencia media de los candidatos esté entre la media de sus logits de las dos pruebas (2,52 y 3,37) ± 2 * ,58.

Las medias de los estadísticos de *Infit* y de *Outfit* de los examinados en ambas pruebas, así como de las de sus desviaciones típicas, revelan que los sujetos se ajustan de forma bastante aceptable al modelo. Ninguna de las medias de los valores de *Infit* y *Outfit* supera ampliamente el valor de 1, que es cuando se manifiesta mayor desajuste del esperado.

El 8,49% de los candidatos del examen D031 y el 6,52% de los candidatos del examen D032 presentan un desajuste severo con las predicciones del modelo. Este porcentaje, que corresponde a los candidatos con estadísticos *Infit* y/o *Outfit* > 2, es bajo.

7.1.5. Atributos

Tal y como se observa en las tablas 40 y 41 no existen grandes diferencias de dificultad entre los atributos evaluados. Esta relativa invarianza podría deberse, como se ha comentado antes, a un efecto de halo. El atributo más fácil en ambos exámenes es la Holística (–,14 *logit* en el examen D031 y –,22 *logit* en el examen D032), esto significa que es en esta variable en la que mayor puntuación obtienen los candidatos; en la que menor puntuación obtienen —la más difícil— es Corrección y alcance (,15 *logit* en el examen D031 y ,33 *logit* en el examen D032). El atributo Adecuación y coherencia se mantiene en una posición intermedia, –,01 *logit* en el examen D031 y –,11 *logit* en el examen D032.

Tabla 40
Valores y estadísticos de los atributos
Modelo de examen D031

Atributo	Promedio	Dificultad	SE	*Infit*	*Outfit*	RiX
Holística	2,05	-,14	,06	,88	,85	,78
Adec/Coh	2,03	-,01	,06	1,15	1,18	,75
Corr/Alc	2,01	,15	,06	,95	1,01	,72
Media	*2,03*	*,00*	*,06*	*,99*	*1,01*	*,75*
DT	*,02*	*,15*	*,00*	*,14*	*,16*	*,03*

ISR (Criterion Separation Reliability) = ,85

Tabla 41
Valores y estadísticos de los atributos
Modelo de examen D032

Atributo	Promedio	Dificultad	SE	*Infit*	*Outfit*	RiX
Holística	2,18	-,22	,07	,94	,97	,73
Adec/Coh	2,16	-,11	,07	1,11	1,08	,71
Corr/Alc	2,08	,33	,07	,95	,99	,72
Media	*2,14*	*,00*	*,07*	*1,00*	*1,01*	*,72*
DT	*,06*	*,29*	*,00*	*,10*	*,06*	*,01*

ISR (Criterion Separation Reliability) = ,94

La precisión de las estimaciones de la dificultad de los atributos es muy alta: el error típico de la medida (SE) es muy pequeño y los estadísticos de fiabilidad global (ISR –Criterion Separation Reliability–) son elevados. También se observa un adecuado ajuste de los atributos al modelo. Son positivas las elevadas correlaciones atributo-escala (promedio de ,75 en el examen D031 y de ,72 en el examen D032), ya que "manifiestan que existe un patrón semejante de competencia en los dominios evaluados, por lo que es adecuado combinarlos en una única puntuación para reflejar el rendimiento de los candidatos." (Prieto 2011: 237).

7.1.5.1. Análisis con el modelo PCM aplicado a la faceta atributo

En las tablas 42 y 43 es posible visualizar los valores de paso del conjunto de calificadores en cada uno de los atributos; en las dos siguientes, en la 44 y en

la 45, se detalla el uso de las puntuaciones en cada uno de ellos, también por el conjunto de examinadores. Se observa, en líneas generales, que el comportamiento de los calificadores en la utilización de las bandas de calificación de los tres atributos es bastante homogéneo, aunque sí es posible detectar algunas tendencias. En las dos pruebas de ambos exámenes se advierte que el atributo en el que el valor de paso entre las calificaciones 0–1 se sitúa en una posición inferior en la escala *logit* es Adecuación y coherencia (–4,93 *logit* en la prueba del examen D031 y –5,59 *logit* en la del examen D032). Sin embargo, en este mismo atributo, el paso entre las puntuaciones 1–2 se sitúa en las dos pruebas en la posición más elevada en la escala *logit* (,19 *logit* en la prueba del examen D031 y ,54 *logit* en la del examen D032) de todas. La consecuencia de estas variaciones es que la amplitud de la calificación 1 en el atributo Adecuación y coherencia es mayor que en las otras dos. Por el contrario, el atributo que presenta una menor amplitud de la puntuación 1, también en las dos pruebas, es la escala Holística (de –3,91 *logit* a –,81 *logit* en la prueba del examen D031 y de –4,26 *logit* a –,53 *logit* en la del D032). En este mismo atributo Holística es donde se localiza antes el paso entre las calificaciones 2 y 3 en ambas pruebas (4,72 *logit* en la del examen D031 y 4,79 *logit* en la del D032).

Tabla 42
Mapa de la variable EE mayo de 2012
Modelo de examen D031

Measr	+Candidato	-Calificador	-Tarea	-Atributo	S.1	S.2	S.3
9	*				(3)	(3)	(3)
	.						
8	*						
	.						
7	.						
	.						
	**						
6	*.						
	*.						
	**.					---	
5	****						
	***.				---		---
	***						
4	*****						
	*******						
	****.						
3	******						
	****.						
	****				2	2	
2	****.						2
	**.	12					
	****.	10 7					
1	***.						
	**.	2					
	**	8 9	Tarea 3		---		
* 0	* **.	* 11 5	* Tarea 2	* Adec-Coh Corr-Alc Holistica	*	*	*
	*	4	Tarea 1				
	**.					---	---
-1	**.	1					
	*.						
	.						
-2	*	3					
	.	6			1		1
	.					1	
-3	.						
	.						
-4	.						---

-5	.				(0)	(0)	(0)
Measr	* = 3	-Calificador	-Tarea	-Atributo	S.1	S.2	S.3

Measr = Medición Rasch (*logit*)
S.1: Model = ?,?,?,1,R3 ; Atributo: Adec-Coh
S.2: Model = ?,?,?,2,R3 ; Atributo: Corr-Alc
S.3: Model = ?,?,?,3,R3 ; Atributo: Holistica

Tabla 43
Mapa de la variable EE mayo de 2012
Modelo de examen D032

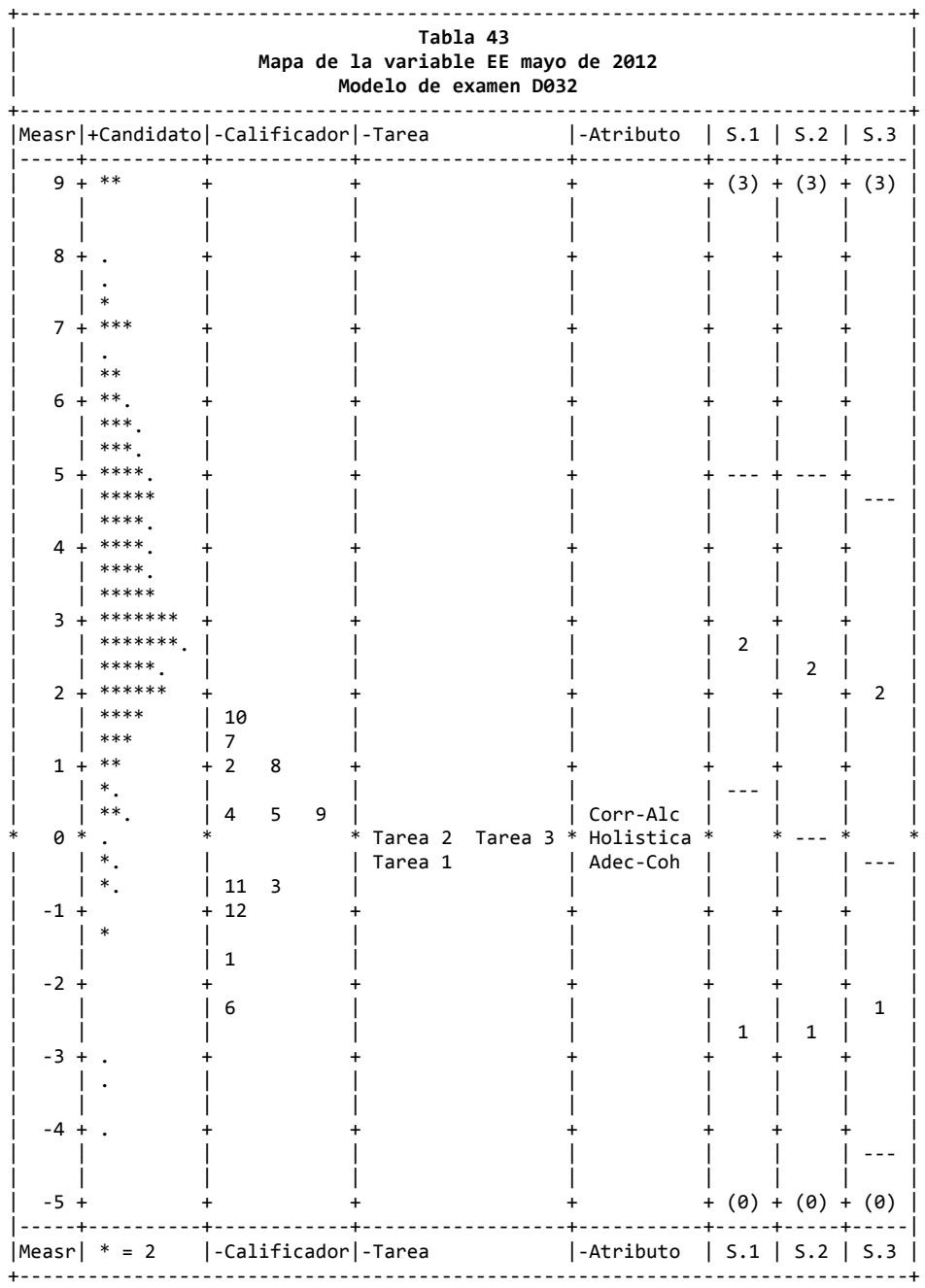


Tabla 44
Uso de las puntuaciones por los calificadores
Modelo de examen D031

Puntuación	Frecuencia absoluta			Frecuencia Relativa			Promedio			*Outfit*			Valor de paso			SE		
	Atributo			Atributo			Atributo			Atributo			Atributo			Atributo		
	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol
0	10	11	25	1%	1%	2%	-1,69	-1,72	-1,66	1,4	1,8	2,1	--	--	--	--	--	--
1	300	228	211	28%	15%	14%	-,16	-,69	-1,13	1,1	1,0	,8	-4,93	-4,76	-3,91	,35	,32	,23
2	824	1003	942	55%	67%	62%	2,56	2,53	2,24	1,1	1,2	,8	,19	-,60	-,81	,09	,10	,10
3	371	261	341	25%	17%	22%	5,35	5,42	5,33	1,0	1,1	,8	4,74	5,36	4,72	,08	,09	,08

Ad-C: Adecuación y coherencia
C-Al: Corrección y alcance
Hol: Holística

Tabla 45
Uso de las puntuaciones 45 los calificadores
Modelo de examen D032

Puntuación	Frecuencia absoluta			Frecuencia Relativa			Promedio			*Outfit*			Valor de paso			SE		
	Atributo			Atributo			Atributo			Atributo			Atributo			Atributo		
	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol
0	3	7	9	0%	1%	1%	-2,25	-4,05	-3,05	1,3	,9	2,1	--	--	--	--	--	--
1	150	146	98	14%	14%	9%	,66	-,02	-,15	1,2	,9	,9	-5,59	-4,99	-4,26	,64	,46	,43
2	597	680	658	55%	63%	61%	3,13	2,77	2,70	,9	1,2	1,1	,54	-,12	-,53	,11	,11	,13
3	328	243	312	30%	23%	29%	5,73	5,33	5,38	1,0	1,0	,9	5,04	5,10	4,79	,09	,09	,09

Ad-C: Adecuación y coherencia
C-Al: Corrección y alcance
Hol: Holística

La combinación en el análisis de un modelo PCM aplicado a la faceta tarea y un modelo RSM aplicado al resto de facetas ya se mostró en el § 7.1.2.2.3. *Análisis de los pasos de las curvas características de las categorías*. En dicho apartado se presentaba el mapa de la variable de un análisis en el que se combinaba un modelo PCM aplicado a la faceta calificador con un modelo RSM aplicado al resto de tareas (tablas 23 y 24).

7.2. Análisis realizado con el modelo PCM aplicado a dos facetas

Por medio de la utilización de este nuevo modelo de análisis en Facets, con la combinación del análisis con el modelo PCM aplicado a dos facetas, es posible obtener información complementaria a la ya obtenida por medio de los análisis anteriormente realizados.

7.2.1. Análisis con el modelo PCM aplicado a las facetas atributo y tarea

En las tablas 46 y 48 es posible visualizar los valores de paso de cada uno de los atributos en cada una de las tareas y en las tablas 47 y 49 se presenta el resumen de los principales estadísticos.

Los valores de paso de cada uno de los atributos en las tres tareas son bastantes homogéneos, tal y como se indicaba en el § 7.1.5. El valor de paso de las puntuaciones 1–2 que se sitúa en una posición más elevada en el mapa de la variable es el del atributo Adecuación y coherencia (significa que a los candidatos les resulta más difícil conseguir un 2 en este atributo que en cualquiera de los otros dos), excepto en la tarea 2 de la prueba del examen D032, en la que el paso entre estas dos calificaciones se localiza en una posición de alrededor de 2 *logit* inferior a los de los otros dos atributos.

Tabla 46
Mapa de la variable EE mayo de 2012
Modelo de examen D031

```
|Measr|+Candidato  |-Calificador|-Tarea   |-Atributo                    | S.1 | S.2 | S.3 | S.4 | S.5 | S.6 | S.7 | S.8 | S.9 |
|-----+------------+------------+---------+-----------------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|   9 + *.        +            +         +                             + (3) + (3) + (3) + (3) + (3) + (3) + (3) + (3) + (3) |
|     |           |            |         |                             |     |     |     |     |     |     |     |     |     |
|     | *         |            |         |                             |     |     |     |     |     |     |     |     |     |
|   8 + *         +            +         +                             +     +     +     +     +     +     +     +     +     |
|     | .         |            |         |                             |     |     |     |     |     |     |     |     |     |
|     |           |            |         |                             |     |     |     |     |     |     |     |     |     |
|   7 + .         +            +         +                             +     +     +     +     +     +     +     +     +     |
|     | **        |            |         |                             |     |     |     |     |     |     |     |     |     |
|     | **        |            |         |                             |     |     |     |     |     |     |     |     |     |
|   6 + **        +            +         +                             +     +     +     +     +     +     +     +     +     |
|     | **        |            |         |                             |     |     |     |     |     |     |     | --- |     |
|     | ****      |            |         |                             |     | --- |     |     | --- |     |     |     |     |
|   5 + *****.    +            +         +                             +     +     +     +     +     + --- + --- +     +     |
|     | *****     |            |         |                             | --- |     |     | --- |     |     |     |     | --- |
|     | ****.     |            |         |                             |     |     | --- |     |     |     |     |     |     |
|   4 + ********  +            +         +                             +     +     +     +     +     +     +     +     +     |
|     | ********* |            |         |                             |     |     |     |     |     |     |     |     |     |
|     | *****.    |            |         |                             |     |     |     |     |     |     |     |     |     |
|   3 + ********** +            +         +                             +     +     +     +     +     +     +     +     +     |
|     | *******   |            |         |                             |     |     |     |     |     |     |     |     |     |
|     | ******    |            |         |                             |  2  |  2  |     |  2  |  2  |     |  2  |  2  |     |
|   2 + ******.   +            +         +                             +     +     +  2  +     +     +  2  +     +     +  2  |
|     | ***       | 12         |         |                             |     |     |     |     |     |     |     |     |     |
|     | ******.   | 10  7      |         |                             |     |     |     |     |     |     |     |     |     |
|   1 + ******.   +            +         +                             +     +     +     +     +     +     +     +     +     |
|     | ***.      | 2          |         |                             |     |     |     |     |     |     |     |     |     |
|     | ***.      | 8   9      | Tarea 3 |                             | --- |     |     |     |     |     |     |     |     |
*   0 * ***.      * 11  5      * Tarea 2 * Adec-Coh   Corr-Alc   Holistica *     *     *     * --- *     *     * --- *     *     *
|     | *.        | 4          | Tarea 1 |                             |     | --- |     |     |     |     |     |     |     |
|     | ****      |            |         |                             |     |     | --- |     | --- | --- |     | --- |     |
|  -1 + ****      + 1          +         +                             +     +     +     +     +     +     +     +     + --- |
|     | **        |            |         |                             |     |     |     |     |     |     |     |     |     |
|     | *         |            |         |                             |     |     |     |     |     |     |     |     |     |
|  -2 + *.        + 3          +         +                             +     +     +     +     +     +     +     +     +     |
|     | *         | 6          |         |                             |  1  |     |  1  |  1  |     |     |  1  |     |  1  |
|     | *         |            |         |                             |     |  1  |     |     |  1  |  1  |     |  1  |     |
|  -3 + .         +            +         +                             +     +     +     +     +     +     +     +     +     |
|     | .         |            |         |                             |     |     |     |     |     |     |     |     |     |
|     |           |            |         |                             |     |     |     |     |     |     |     |     |     |
|  -4 + *         +            +         +                             +     +     + --- +     +     +     +     +     + --- |
|     |           |            |         |                             |     |     |     |     |     | --- |     |     |     |
|     |           |            |         |                             |     | --- |     |     | --- |     |     |     |     |
|  -5 + .         +            +         +                             + (0) + (0) + (0) + (0) + (0) + (0) + (0) + (0) + (0) |
|-----+------------+------------+---------+-----------------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----|
|Measr| * = 2      |-Calificador|-Tarea   |-Atributo                    | S.1 | S.2 | S.3 | S.4 | S.5 | S.6 | S.7 | S.8 | S.9 |
```

S.1: Model = ?,?,1,1,R3 ; Tarea: Tarea 1 Atributo: Adec-Coh
S.2: Model = ?,?,1,2,R3 ; Tarea: Tarea 1 Atributo: Corr-Alc
S.3: Model = ?,?,1,3,R3 ; Tarea: Tarea 1 Atributo: Holistica
S.4: Model = ?,?,2,1,R3 ; Tarea: Tarea 2 Atributo: Adec-Coh
S.5: Model = ?,?,2,2,R3 ; Tarea: Tarea 2 Atributo: Corr-Alc
S.6: Model = ?,?,2,3,R3 ; Tarea: Tarea 2 Atributo: Holistica
S.7: Model = ?,?,3,1,R3 ; Tarea: Tarea 3 Atributo: Adec-Coh
S.8: Model = ?,?,3,2,R3 ; Tarea: Tarea 3 Atributo: Corr-Alc
S.9: Model = ?,?,3,3,R3 ; Tarea: Tarea 3 Atributo: Holistica

Tabla 47
Uso de las puntuaciones por los calificadores
Modelo de examen D031

	Puntuación	Frecuencia absoluta			Frecuencia Relativa			Promedio			*Outfit*			Valor de paso			SE		
		Atributo			Atributo			Atributo			Atributo			Atributo			Atributo		
		Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol
Tarea 1	0	3	3	8	1%	1%	2%	-,99	-1,70	-2,13	1,5	1,4	1,3	--	--	--	--	--	--
	1	101	72	70	20%	14%	14%	,.04	-,64	-,87	1,1	,9	,8	-4,86	-4,79	-3,81	,65	,62	,41
	2	252	320	287	50%	64%	57%	2,71	2,61	2,32	1,1	,9	,7	,39	-,42	-,60	,16	,17	,17
	3	146	107	142	29%	21%	28%	5,29	5,45	5,24	1,0	1,2	,9	4,47	5,21	4,41	,13	,15	,13
Tarea 2	0	3	4	5	1%	1%	1%	-,81	-,89	-,01	1,8	3,2	3,9	--	--	--	--	--	--
	1	96	74	71	19%	15%	14%	-,22	-,55	-1,15*	1,1	1,2	,6	-4,96	-4,66	-4,33	,62	,56	,47
	2	278	341	324	55%	68%	64%	2,53	2,55	2,34	,9	1,2	,7	,17	-,74	-,70	,16	,17	,17
	3	127	84	106	25%	17%	21%	5,50	5,56	5,57	,9	1,0	,8	4,80	5,40	5,03	,14	,15	,15
Tarea 3	0	4	4	12	1%	1%	2%	-2,89	-2,58	-2,04	,9	,9	1,9	--	--	--	--	--	--
	1	103	82	70	21%	16%	14%	-,34	-,88	-1,34	1,2	1,0	,8	-4,95	-4,85	-3,72	-4,95	,52	,34
	2	294	342	331	59%	69%	65%	2,42	2,40	2,09	1,3	1,4	,9	-,01	-,67	-1,09	-,01	,16	,17
	3	98	70	93	20%	14%	18%	5,27	5,24	5,31	1,0	1,2	,7	4,96	5,51	4,81	4,96	,17	,15

Ad-C: Adecuación y coherencia
C-Al: Corrección y alcance
Hol: Holística

Es destacable la incongruencia que se observa en los promedios del atributo Holística en la tarea 2 de la prueba del examen D031. El promedio de la puntuación 1 es 1,14 *logit* inferior al de la calificación 0. Este hecho no es deseable, aunque el bajo porcentaje de alumnos a los que se les asignó la puntuación de 0 puede tener relación con este hecho.

Tabla 48
Mapa de la variable EE mayo de 2012
Modelo de examen D032

Measr	+Candidato	-Calificador	-Tarea	-Atributo	S.1	S.2	S.3	S.4	S.5	S.6	S.7	S.8	S.9
8	**.				(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)
	.												
7	**												
	**												
	.												
6	*.												
	***												
	**					---							
5	****.								---	---		---	
	****				---		---				---		---
	******												
4	***.												
	*****												
	*****												
3	****.												
	*******				2	2							
	********.							---	2	2	2	2	
2	*****						2						2
	******	10											
	****.	7											
1	**.	2 8											
	**		Tarea 2		---								
	*.	4 5 9		Adec-Coh									
0	**.		Tarea 3	Corr-Alc		---		2	---		---		
	.			Holistica			---			---		---	
	*.	11 3	Tarea 1										---
-1	*.	12											
	*	1											
-2													
		6			1		1	---		1	1	1	1
						1			1				
-3													
	.												
	.												
-4													---
	.						---			---			
											---	---	
-5					(0)	(0)	(0)	(1)	(0)	(0)	(0)	(0)	(0)

Measr	* = 2	-Calificador	-Tarea	-Atributo	S.1	S.2	S.3	S.4	S.5	S.6	S.7	S.8	S.9

S.1: Model = ?,?,1,1,R3 ; Tarea: Tarea 1 Atributo: Adec-Coh
S.2: Model = ?,?,1,2,R3 ; Tarea: Tarea 1 Atributo: Corr-Alc
S.3: Model = ?,?,1,3,R3 ; Tarea: Tarea 1 Atributo: Holistica
S.4: Model = ?,?,2,1,R3 ; Tarea: Tarea 2 Atributo: Adec-Coh
S.5: Model = ?,?,2,2,R3 ; Tarea: Tarea 2 Atributo: Corr-Alc
S.6: Model = ?,?,2,3,R3 ; Tarea: Tarea 2 Atributo: Holistica
S.7: Model = ?,?,3,1,R3 ; Tarea: Tarea 3 Atributo: Adec-Coh
S.8: Model = ?,?,3,2,R3 ; Tarea: Tarea 3 Atributo: Corr-Alc
S.9: Model = ?,?,3,3,R3 ; Tarea: Tarea 3 Atributo: Holistica

Tabla 49
Uso de las puntuaciones por los calificadores
Modelo de examen D032

	Puntuación	Frecuencia absoluta			Frecuencia Relativa			Promedio			*Outfit*			Valor de paso			SE		
		Atributo			Atributo			Atributo			Atributo			Atributo			Atributo		
		Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol
Tarea 1	0	1	1	3	0%	0%	1%	-3,68	-4,52	-2,80	,8	,5	1,9	--	--	--	--	--	--
	1	54	37	32	15%	10%	9%	,30	,12	,51	,8	,9	1,5	-5,47	-5,43	-4,32	,94	1,11	,88
	2	186	227	210	52%	63%	58%	2,72	3,27	3,21	,9	1,6	1,7	,70	,01	-,41	,18	,22	,26
	3	118	94	115	33%	26%	32%	5,19	5,76	5,90	1,4	1,0	,6	4,77	5,42	4,73	,16	,16	,15
Tarea 2	0	--	3	2	--	1%	1%	--	-4,59	-5,34	--	,6	,1	--	--	--	--	--	--
	1	51	59	32	14%	16%	9%	-,98	-,45	-1,04	3,2	,8	,4	--	-4,91	-4,41	--	-4,91	,62
	2	208	230	228	58%	64%	63%	1,32	2,15	1,95	1,4	1,0	,7	-2,38	-,15	-,54	,26	-,15	,19
	3	101	68	98	28%	19%	27%	4,11	4,99	4,80	,4	1,2	1,7	2,38	5,07	4,95	,14	5,07	,17
Tarea 3	0	2	3	4	1%	1%	1%	-2,80	-3,36	-1,73	,7	1,3	4,8	--	--	--	--	--	--
	1	45	50	34	13%	14%	10%	-,51	,25	,06	,6	1,2	1,2	-4,69	-4,66	-3,95	,67	,82	,76
	2	203	223	220	57%	62%	62%	2,17	2,91	2,87	,9	1,4	1,5	,03	-,23	-,67	,18	,20	,24
	3	109	81	99	30%	23%	28%	4,74	5,38	5,43	1,6	,8	,7	4,66	4,89	4,62	,16	,16	,15

Ad-C: Adecuación y coherencia
C-Al: Corrección y alcance
Hol: Holística

7.2.2. Análisis con el modelo PCM aplicado a las facetas calificador y tarea

En los mapas mostrados en las tablas 23 y 24 y en el resumen de las mismas, 25 y 26, se analizaba el comportamiento del equipo de calificadores en cada una de las tareas de las dos pruebas. Pero aquel análisis, aunque eficaz para el estudio de la tendencia central, no permite saber si cada uno de los examinadores ha calificado de igual modo cada una de las tareas o si su proceder ha variado dependiendo de la tarea que estuviera puntuando. En las tablas 27 y 28, así como en los resúmenes presentados en las tablas 35 y 36 —integradas también en el proceso de análisis de la tendencia central—, en las que se visualizaba la localización de los valores de paso de los calificadores en el conjunto de tareas, ya se observaba que el comportamiento de cada uno de los calificadores en las tres tareas agrupadas presentaba diferencias considerables en la localización de los valores de paso.

Pero para conocer cuál ha sido el comportamiento de los calificadores individualmente en cada una de las tareas, es posible realizar una nueva modificación el modelo de análisis del programa FACETS y combinar un modelo PCM aplicado a las facetas calificador y tarea con un modelo RSM aplicado al resto de facetas.

Este nuevo procedimiento de estudio de cada uno de los calificadores posibilita, por un lado, ampliar el foco sobre la manera que tienen los calificadores de aplicar las bandas de puntuación y, por otro, permite analizar la severidad y los valores de paso de las puntuaciones en cada una de las tareas para cada uno de los calificadores.

7.2.2.1. Análisis de la utilización de las bandas de calificación en cada una de las tareas

En las tablas 50 y 52 se analiza el comportamiento de cada uno de los calificadores en las tres tareas de manera independiente. En el análisis realizado en la tabla 27 (y en su resumen en la tabla 17) se indicaba que los promedios de las puntuaciones 0 y 1 del calificador 4 se encontraban invertidos: el promedio de la banda de calificación más baja, el 0, era superior a la de la que se encontraba en la posición inmediatamente superior, el 1. Por medio de este nuevo análisis es posible detectar si este hecho sucede en todas las tareas o únicamente en alguna de ellas. Pues bien, en la tabla resumen del mapa de la variable de la prueba del examen D031 (tabla 50) se constata que la inversión de los promedios de las puntuaciones 0 y 1 en este calificador se producen en las tareas 1 y 2, mientras que en la tarea 3 los promedios son los esperables: los de las puntuaciones más bajas son inferiores a los de las calificaciones más altas. En relación con este mismo calificador, también se indicaba que el valor de paso de las calificaciones adyacentes 0–1 se localizaban en una posición superior en el mapa de la variable al de las

puntuaciones 1–2. Pues bien, en este análisis se constata que dicha incongruencia se produce en las tres tareas; en todas ellas el valor de paso de las calificaciones 0–1 se localiza en una posición más elevada en el mapa de la variable que el de 1–2. Estos datos parecen indicar que el calificador 4 tiene en general problemas para asignar las dos bandas de calificación inferior.

También se había indicado que el calificador 3 era el único en el que todos los promedios de sus calificaciones se localizaban en la parte positiva de la escala de *logit*. Si se analizan los promedios de las puntuaciones 0 y 1 en la tarea 3 de este calificador se constata que es superior el promedio de la calificación 0 que el de la puntuación 1. Esta misma discordancia en los promedios de las calificaciones de las puntuaciones 0 y 1 se observa en la tarea 2 del calificador 8 y en la tarea 3 del calificador 11. No obstante, es preciso destacar que el porcentaje de utilización de la banda de calificación 0 es muy bajo en todos los casos, lo que dificulta la posibilidad de extraer conclusiones relevante de este análisis.

En el análisis de la prueba del examen D032 (tabla 28) se había destacado que el único calificador en el que el promedio de la banda de calificación más baja —el 0— era superior al de la puntuación 1 era el 3. En las tablas 51 y 53 se constata que esta incongruencia sucede únicamente en la tarea 3, mientras que en las otras dos la ordenación es coherente. Es conveniente tener en cuenta el hecho de que este calificador ha utilizado en una única ocasión la puntuación 0, por lo que los resultados no son concluyentes.

El calificador 4, que en la prueba anteriormente analizada mostraba incongruencias en los promedios de las calificaciones 0 y 1 en dos de las tareas, en esta prueba no los presenta ya que no utiliza en ninguna ocasión las puntuaciones 0 o 1. Este es el motivo por el que el programa FACETS no proporciona información acerca del valor de paso de la calificación 2 en la tarea 1 de este calificador. El programa estadístico no facilita datos de los valores de paso entre calificaciones adyacentes si el calificador utiliza

únicamente una o dos de las cuatro puntuaciones disponibles (de 0 a 3). En la tarea 3 sí utiliza en una ocasión la calificación 1, aunque el promedio de esta puntuación similar al de la puntuación 2.

Se encuentra la misma discordancia en los promedios de las calificaciones de las puntuaciones 0 y 1 en la tarea 2 del calificador 6 y en la tarea 3 del calificador 10. No obstante, es preciso destacar que en ambos casos el porcentaje de utilización de la banda de calificación 0, al igual que sucedía en la prueba anteriormente analizada, es muy bajo en todos los casos, lo que dificulta la posibilidad de sacar conclusiones relevantes.

Tabla 50 (1 de 3)
Mapa de la variable EE mayo de 2012
***Logit* + Valores de paso de los calificadores**
Modelo de examen D031

```
+-----------------------------------------------------------------------+
|Measr|+Candidato|-Calificador|-Tarea   |-Atributo                      |
|-----+----------+------------+---------+-------------------------------+
|   8 + *        +            +         +                               +
|     | .        |            |         |                               |
|     |          |            |         |                               |
|   7 + .        +            +         +                               +
|     | .        |            |         |                               |
|     | .        |            |         |                               |
|   6 + *.       +            +         +                               +
|     | .        |            |         |                               |
|     | *        |            |         |                               |
|   5 + *.       +            +         +                               +
|     | *.       |            |         |                               |
|     | **       |            |         |                               |
|   4 + **.      +            +         +                               +
|     | ***.     |            |         |                               |
|     | **.      |            |         |                               |
|   3 + **.      +            +         +                               +
|     | ****.    | 7          |         |                               |
|     | ******** |            |         |                               |
|   2 + *****.   + 12         +         +                               +
|     | ****     |            |         |                               |
|     | ***.     |            |         |                               |
|   1 + *****    + 10  2      +         +                               +
|     | **.      |            |         |                               |
|     | ****     | 5          | Tarea 2 |                               |
*   0 * **.      * 11         * Tarea 3 * Adec-Coh   Corr-Alc   Holistica *
|     | *****.   | 4          | Tarea 1 |                               |
|     | **.      | 8          |         |                               |
|  -1 + **       + 3   9      +         +                               +
|     | **       | 1          |         |                               |
|     | ***      |            |         |                               |
|  -2 + ***      +            +         +                               +
|     | *.       |            |         |                               |
|     | *        |            |         |                               |
|  -3 + *.       + 6          +         +                               +
|     | *.       |            |         |                               |
|     | .        |            |         |                               |
|  -4 +          +            +         +                               +
|     |          |            |         |                               |
|     |          |            |         |                               |
|  -5 +          +            +         +                               +
|     |          |            |         |                               |
|     | .        |            |         |                               |
|  -6 +          +            +         +                               +
|     |          |            |         |                               |
|     |          |            |         |                               |
|  -7 +          +            +         +                               +
|     |          |            |         |                               |
|     | .        |            |         |                               |
|  -8 + .        +            +         +                               +
|-----+----------+------------+---------+-------------------------------+
|Measr| * = 3    |-Calificador|-Tarea   |-Atributo                      |
+-----------------------------------------------------------------------+
 Measr = Medición Rasch (logit)
```

Tabla 50 (2 de 3)
Mapa de la variable EE mayo de 2012
***Logit* + Valores de paso de los calificadores**
Modelo de examen D031
Calificadores 1-6

Measr	1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	6.1	6.2	6.3
8	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)
7																		
6																		
5																		
4																		
3																		
2																		
1																		
0																		
-1																		
-2																		
-3																		
-4																		
-5																		
-6																		
-7																		
-8	(0)	(1)	(0)	(1)	(1)	(0)	(1)	(1)	(0)	(0)	(0)	(0)	(0)	(1)	(1)	(1)	(0)	(0)
Measr	1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	6.1	6.2	6.3

Measr = Medición Rasch (*Logit*)
1.1, 1.2, 1.3, 2.1 = Calificador 1, tarea 1; Calificador 1, tarea 2; Calificador 1, tarea 3; Calificador 2, tarea 1…

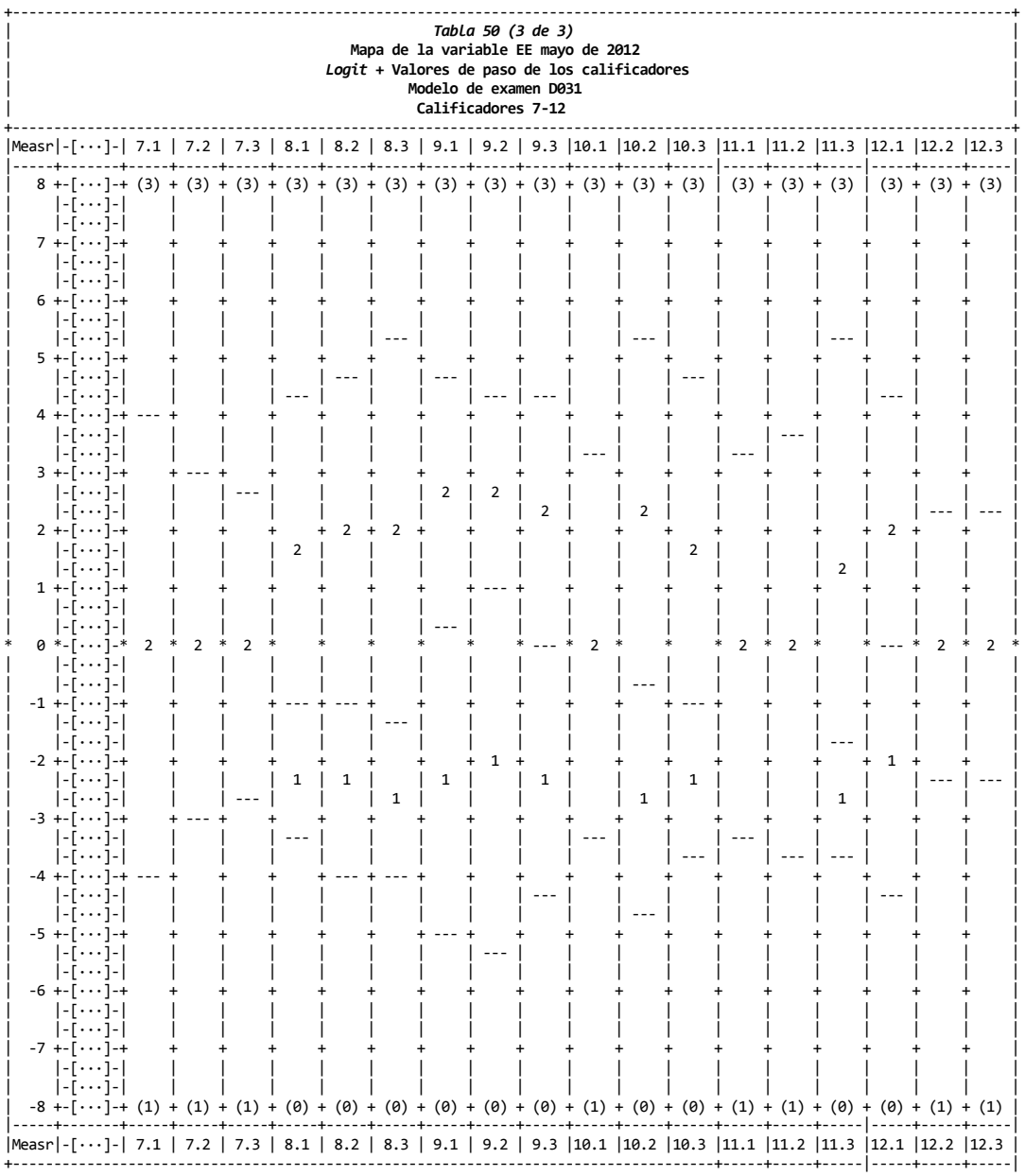


Tabla 50 (3 de 3)
Mapa de la variable EE mayo de 2012
***Logit* + Valores de paso de los calificadores**
Modelo de examen D031
Calificadores 7-12

Measr	7.1	7.2	7.3	8.1	8.2	8.3	9.1	9.2	9.3	10.1	10.2	10.3	11.1	11.2	11.3	12.1	12.2	12.3
8	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)
7																		
6																		
5																		
4																		
3																		
2																		
1																		
0	2	2	2	*	*	*	*	*	*	2	*	*	2	2	*	*	2	2
-1																		
-2				1	1		1		1			1						
-3						1					1				1			
-4																		
-5																		
-6																		
-7																		
-8	(1)	(1)	(1)	(0)	(0)	(0)	(0)	(0)	(0)	(1)	(0)	(0)	(1)	(1)	(0)	(0)	(1)	(1)
Measr	7.1	7.2	7.3	8.1	8.2	8.3	9.1	9.2	9.3	10.1	10.2	10.3	11.1	11.2	11.3	12.1	12.2	12.3

Measr = Medición Rasch (*Logit*)
7.1, 7.2, 7.3, 8.1 = Calificador 7, tarea 1; Calificador 7, tarea 2; Calificador 7, tarea 3; Calificador 8, tarea 1…

Tabla 51
Uso de las puntuaciones por los calificadores
Modelo de examen D031

	Puntuación	Frecuencia absoluta			Frecuencia Relativa			Promedio			*Outfit*			Valor de paso			SE		
		Tarea			Tarea			Tarea			Tarea			Tarea			Tarea		
		1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Calif. 1	0	1	--	1	1%	--	1%	-3,85	--	-4,11	,1	--	,1	--	--	--	--	--	--
	1	12	15	17	9%	12%	13%	-,49	-1,91	-,47	,6	2,2	,5	-3,49	--	-3,96	,92	--	,94
	2	65	76	68	51%	59%	54%	1,76	1,46	1,75	,4	1,1	,7	-,54	-2,66	-,28	,33	,48	,30
	3	49	38	41	39%	29%	32%	5,02	5,05	5,14	1,1	,2	1,0	4,02	2,66	4,25	,27	,25	,28
Calif. 2	0	--	--	2	--		1%	--	--	-3,09	--	--	,2	--	--	--	--	--	--
	1	24	16	15	17%	11%	11%	-1,70	--	-2,92	2,3	,7	,1	--	--	-3,74	--	--	,47
	2	99	115	109	70%	82%	78%	,82	-3,72	,56	1,5	,7	1,3	-2,83	-3,72	-1,56	,34	,36	,27
	3	18	10	14	13%	7%	10%	3,23	3,72	3,85	,6	,3	3,2	2,83	3,72	5,29	,26	,36	,51
Calif. 3	0	--	--	1	--	--	1%	--	--	-,69	--	--	,5	--	--	--	--	--	--
	1	15	23	7	12%	18%	6%	-,58	-,63	-1,02*	1,2	1,8	,3	--	--	-2,45	--	--	,66
	2	33	47	64	26%	36%	51%	1,01	,64	1,03	1,4	,9	,6	-,94	-1,35	-1,27	,36	,33	,33
	3	81	59	54	63%	46%	43%	3,56	3,88	4,04	1,0	,3	1,6	,94	1,35	3,72	,27	,26	,26
Calif. 4	0	3	6	3	2%	4%	2%	-,77	-,34	-2,32	,8	6,4	,2	--	--	--	--	--	--
	1	4	2	1	3%	1%	1%	-1,65*	-1,33*	,19	,0	,7	2,2	-,91	-,22	-,34	,64	,61	,80
	2	77	107	118	57%	79%	87%	1,53	1,47	1,65	,8	1,3	,8	-2,02	-3,62	-4,37	,44	,52	,68
	3	51	20	13	38%	15%	10%	3,22	2,45	3,90	1,1	1,5	,8	2,93	3,84	4,71	,22	,27	,32
Calif. 5	0	3	--	--	2%	--	--	-3,21	--	--	,2	--	--	--	--	--	--	--	--
	1	21	17	37	16%	13%	29%	-1,95	-2,65	-1,60	,3	1,0	3,5	-3,43	--	--	,44	--	--
	2	65	79	73	50%	63%	57%	1,04	,59	1,49	1,5	1,2	2,9	-,23	-2,73	-2,49	,28	,35	,34
	3	40	30	18	31%	24%	14%	3,06	2,61	3,35	2,5	1,3	,3	3,65	2,73	2,49	,29	,26	,25
Calif. 6	0	--	1	1	--	1%	1%	--	-4,72	-4,45	--	,1	1,1	--	--	--	--	--	--
	1	1	2	4	1%	3%	5%	-4,51	-,34	,59	,2	,3	1,1	--	-2,63	-4,23	--	2,13	1,92
	2	29	32	44	37%	42%	58%	2,67	3,01	3,86	,6	1,6	3,4	-4,21	-1,09	-,58	1,11	,80	,72
	3	48	41	27	62%	54%	36%	5,44	5,16	5,76	3,2	1,2	1,0	4,21	3,72	4,82	,30	,30	,30
Calif. 7	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	35	27	30	26%	20%	22%	-2,66	-3,79	-3,28	4,1	,6	,9	--	--	--	--	--	--
	2	98	97	91	73%	72%	67%	-,11	-,85	-,54	,7	1,2	1,1	-3,94	-2,96	-2,70	,33	,26	,26
	3	2	11	14	1%	8%	10%	1,73	1,16	,86	,4	1,8	1,7	3,94	2,96	2,70	,47	,41	,35
Calif. 8	0	3	2	4	2%	1%	3%	-2,17	-,80	-1,71	,8	3,1	2,2	--	--	--	--	--	--
	1	16	21	23	12%	15%	17%	-,05	-1,02*	-,60	1,4	,8	1,7	-3,11	-3,83	-3,84	,71	,59	,70
	2	96	101	104	71%	72%	75%	1,79	1,16	1,89	1,1	1,3	,8	-1,28	-,98	-1,36	,33	,27	,30
	3	29	16	8	15%	11%	6%	4,06	4,01	3,36	,7	,9	1,2	4,39	4,81	5,21	,27	,34	,35
Calif. 9	0	2	2	3	1%	1%	2%	-2,36	-3,33	-1,85	1,7	,9	1,6	--	--	--	--	--	--
	1	39	47	42	29%	35%	32%	-1,14	-1,17	-,62	,6	,7	1,8	-5,01	-5,28	-4,41	,79	,70	,72
	2	62	51	61	46%	38%	46%	2,59	1,97	2,48	,7	1,0	,8	,28	,96	-,04	,29	,29	,29
	3	32	35	26	24%	26%	20%	5,53	5,09	5,68	,9	1,0	,4	4,73	4,32	4,45	,29	,30	,29
Calif. 10	0	--	1	2	--	1%	2%	--	-8,79	-5,80	--	,0	,1	--	--	--	--	--	--
	1	33	22	16	30%	20%	15%	-1,81	-1,98	.1,41	6,9	,1	,4	--	-4,79	-3,55	--	,73	,60
	2	72	73	71	65%	67%	66%	1,88	,80	,75	2,7	1,5	1,1	-3,29	-,58	-1,10	,46	,27	,29
	3	6	13	18	5%	12%	17%	4,45	4,28	4,34	,1	2,2	1,3	3,29	5,37	4,64	,31	,49	,39
Calif. 11	0	--	--	3	--	--	2%	--	--	-2,60	--	--	,3	--	--	--	--	--	--
	1	12	8	8	9%	6%	6%	-1,61	-2,88	-2,67*	2,4	,7	,2	--	--	-2,90	--	--	,49
	2	94	95	101	71%	72%	77%	1,21	,72	1,23	1,5	1,4	1,0	-3,48	-3,74	-2,53	,44	,41	,34
	3	26	29	20	20%	22%	15%	4,73	3,52	4,42	,3	1,6	2,3	3,48	3,74	5,43	,26	,28	,35
Calif. 12	0	2	--	--	2%	--	--	-3,75	--	--	,1	--	--	--	--	--	--	--	--
	1	31	41	55	25%	33%	44%	-2,47	-2,97	-2,22	,2	1,1	3,1	-4,31	--	--	,42	--	--
	2	69	70	63	56%	56%	50%	,44	,16	,86	2,6	,7	,8	,05	-2,35	-2,45	,25	,28	,28
	3	22	15	8	18%	12%	6%	2,56	2,57	3,68	4,5	,6	,1	4,26	2,35	2,45	,46	,32	,30

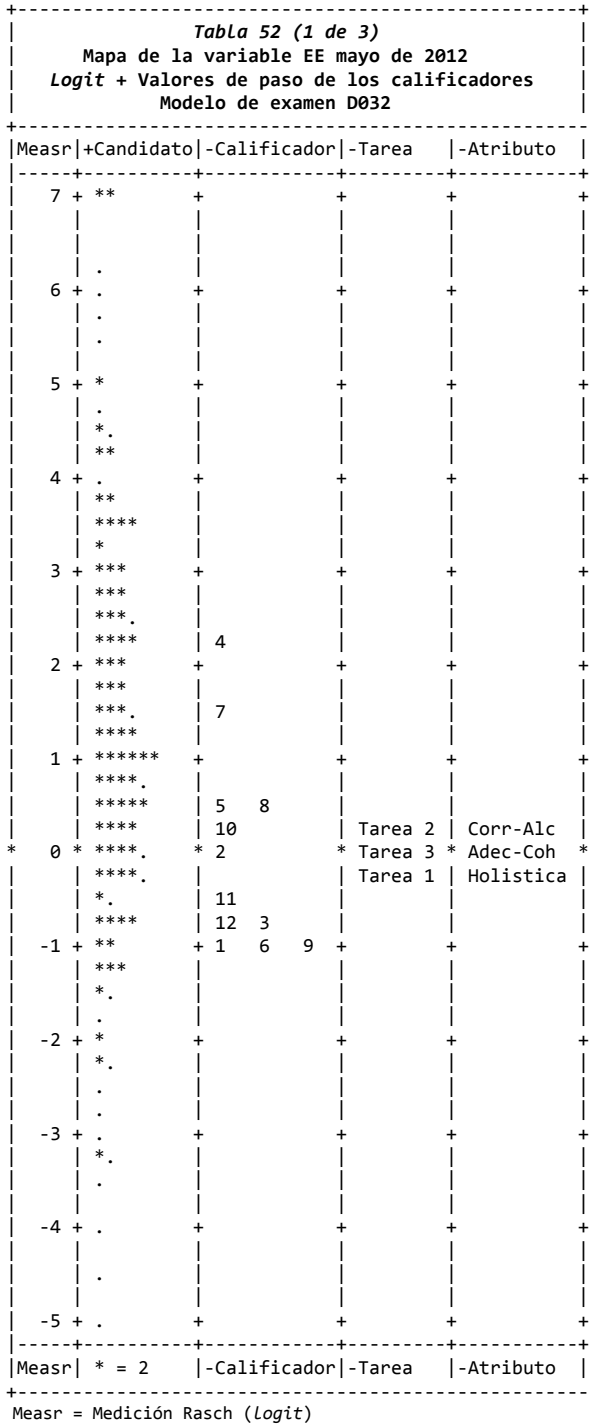


Tabla 52 (1 de 3)
Mapa de la variable EE mayo de 2012
Logit + Valores de paso de los calificadores
Modelo de examen D032
Measr | +Candidato | -Calificador | -Tarea | -Atributo
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
4
7
5 8
10
2
11
12 3
1 6 9
Tarea 2
Tarea 3
Tarea 1
Corr-Alc
Adec-Coh
Holistica
* = 2
Measr = Medición Rasch (logit)

Tabla 52 (2 de 3)
Mapa de la variable EE mayo de 2012
***Logit* + Valores de paso de los calificadores**
Modelo de examen D032
Calificadores 1-6

Measr	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3	4.2	5.1	5.2	5.3	6.2	6.3
7	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)
6														
5				---										
4									---					
3														
2														---
1														
0	2	2			2	2	2	2		2	2	2	2	2
-1														
-2														---
-3	---	---						---						
-4									---					
-5	(1)	(1)	(0)	(0)	(1)	(1)	(1)	(0)	(1)	(1)	(0)	(1)	(1)	(1)

Measr = Medición Rasch (*Logit*)
4.1, 4.2, 4.3, 2.1 = Calificador 4, tarea 1; Calificador 4, tarea 2; Calificador 4, tarea 3; Calificador 2, tarea 1…

Tabla 52 (3 de 3)
Mapa de la variable EE mayo de 2012
***Logit* + Valores de paso de los calificadores**
Modelo de examen D032
Calificadores 7-12

Measr	7.1	7.2	7.3	8.1	8.2	8.3	9.1	9.2	9.3	10.2	10.3	11.1	11.2	11.3	12.1	12.2	12.3
7	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)
6																	
5											---						
4										---							
3												---			---		
2	---		---	2			2				2					---	---
1								2	2								
0	2	2	2	---	2	2				---		2	2	2	2	2	2
-1								1	1		---						
-2	---	---	---	1			1				1					---	---
-3					---	---				---		---	---	---	---		
-4				---			---	---		---							
-5	(1)	(1)	(1)	(0)	(1)	(1)	(0)	(0)	(0)	(1)	(0)	(1)	(1)	(1)	(1)	(1)	(1)
Measr	7.1	7.2	7.3	8.1	8.2	8.3	9.1	9.2	9.3	10.2	10.3	11.1	11.2	11.3	12.1	12.2	12.3

Measr = Medición Rasch (*Logit*)
7.2, 7.3, 8.1 = Calificador 7, tarea 2; Calificador 7, tarea 3; Calificador 8, tarea 1…

Tabla 53
Uso de las puntuaciones por los calificadores
Modelo de examen D032

	Puntuación	Frecuencia absoluta			Frecuencia Relativa			Promedio			*Outfit*			Valor de paso			SE		
		Tarea			Tarea			Tarea			Tarea			Tarea			Tarea		
		1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Calif. 1	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	--	3	1	--	3%	1%	--	-2,05	-,47	--	,4	,5	--	--	--	--	--	--
	2	49	53	40	56%	61%	46%	,88	,73	,36	9,9	,6	1,0	--	-2,94	-3,03	--	,57	,60
	3	38	31	46	44%	36%	53%	3,17	3,29	2,75	,2	1,1	2,8	--	2,94	3,03	--	,30	,30
Calif. 2	0	2	2	--	3%	3%	--	-4,37	-4,70	--	,1	,1	--	--	--	--	--	--	--
	1	11	8	13	15%	11%	17%	-,88	-1,84	-1,80	,7	,2	2,3	-3,40	-3,28	--	,81	,79	--
	2	56	62	55	75%	83%	73%	1,11	,83	,84	1,1	1,0	1,0	-1,04	-1,73	-2,79	,35	,37	,51
	3	6	3	7	8%	4%	9%	2,78	2,77	2,51	1,2	1,2	,3	4,43	5,01	2,79	4,43	,72	,34
Calif. 3	0	--	--	1	--	--	1%	--		,18	--	--	,9	--	--	--	--	--	--
	1	15	17	9	17%	19%	10%	-1,13	-,68	-1,49	1,6	2,5	,3	--	--	-2,84	--	--	,59
	2	47	49	45	52%	54%	51%	,94	,64	,53	,7	,9	1,3	-1,87	-1,94	-,61	,42	,39	,32
	3	28	24	33	31%	27%	38%	3,66	3,32	2,89	,4	,5	2,9	1,87	1,94	3,44	,28	,29	,35
Calif. 4	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	--	1	--	--	1%	--	--	-1,32	--	--	,3	--	--	--	--	--	--	--
	2	60	68	78	63%	71%	81%	-,98	-1,29	-,82	1,4	1,0	5,6	--	-3,92	--	--	,39	--
	3	36	27	18	38%	28%	19%	1,06	1,34	,97	1,0	9,9	,2	--	3,92	--	--	,55	--
Calif. 5	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	10	15	12	10%	15%	12%	-1,76	-2,41	-2,14	,9	,9	,9	--	--	--	--	--	--
	2	59	65	63	59%	66%	64%	,22	,39	,11	1,1	,8	,6	-2,56	-2,73	-2,68	,37	,36	,36
	3	30	19	24	30%	19%	24%	2,96	3,27	3,20	1,2	,6	,7	2,56	2,73	2,68	,30	,31	,31
Calif. 6	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	--	1	2	--	2%	4%	--	1,44	-1,54	--	,8	,3	--	--	--	--	--	--
	2	21	24	22	37%	42%	42%	1,21	.68*	1,26	5,2	,4	1,6	--	-2,47	-1,95	--	,80	,71
	3	36	32	33	63%	56%	58%	3,19	3,24	3,05	,4	1,3	1,2	--	2,47	1,95	--	,34	,34
Calif. 7	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	20	27	20	17%	23%	17%	-1,33	-1,48	-1,62	1,2	1,3	,7	--	--	--	--	--	--
	2	73	76	74	62%	65%	63%	-,01	-,17	-,23	,9	1,4	1,4	-1,94	-2,13	-1,99	,28	,27	,26
	3	24	14	23	24%	12%	23%	1,59	1,69	1,27	1,0	,7	1,4	1,94	2,13	1,99	,28	,30	,29
Calif. 8	0	1	--	--	1%	--	--	-3,51	--	--	,2	--	--	--	--	--	--	--	--
	1	23	21	28	23%	21%	27%	-,72	-2,34	-1,71	1,1	1,5	2,8	-4,40	--	--	,62	--	--
	2	59	68	64	59%	67%	63%	,13	,10	,47	2,3	,9	1,8	-,18	-2,76	-2,65	,26	,34	,34
	3	17	13	10	17%	13%	10%	3,48	3,67	2,97	3,0	,2	,4	4,58	2,76	2,65	,47	,36	,35
Calif. 9	0	2	3	6	2%	3%	6%	-2,45	-3,23	-2,89	1,0	,4	1,1	--	--	--	--	--	--
	1	19	17	22	19%	17%	22%	-,60	-,85	-,20	,8	,7	1,0	-3,88	-3,35	-3,24	,76	,62	,62
	2	41	38	40	41%	38%	40%	2,14	1,74	2,40	1,1	,6	1,2	,36	,24	,10	,35	,34	,35
	3	37	41	31	37%	41%	31%	4,65	4,11	4,61	,8	1,3	,5	3,52	3,11	3,14	,29	,29	,29
Calif. 10	0	--	--	1	--	--	1%	--	--	-1,00	--	--	,9	--	--	--	--	--	--
	1	14	16	9	19%	22%	13%	-1,22	-1,37	-2,25*	,2	9,6	,2	--	--	-3,59	--	--	,77
	2	58	55	51	81%	76%	73%	1,79	1,47	1,20	1,3	1,1	,9	--	-4,10	-1,06	--	-,71	,37
	3	--	1	9	--	1%	13%	--	3,99	3,35	--	,0	1,5	--	4,10	4,65	--	,41	,48
Calif. 11	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	4	5	3	4%	6%	3%	-1,24	-,62	-2,15	1,2	1,3	,5	--	--	--	--	--	--
	2	61	60	60	68%	67%	67%	1,24	,75	,75	1,6	,8	,7	-2,99	-2,85	-3,15	,61	,52	,58
	3	25	25	26	28%	28%	29%	3,29	3,24	3,42	1,0	1,2	1,1	2,99	2,85	3,15	,30	,31	,32
Calif. 12	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	1	5	7	1%	6%	8%	-,64	-1,46	-,92	,7	,9	2,0	--	--	--	--	--	--
	2	39	48	54	43%	53%	60%	1,26	1,34	1,46	,6	,8	,9	-2,92	-2,39	-2,47	,73	-2,39	,60
	3	50	37	29	56%	41%	32%	3,25	3,35	3,95	2,3	1,3	,5	2,92	2,39	2,47	,27	2,39	,27

7.2.2.2. Análisis de los valores de paso de los calificadores en cada una de las tareas.

Por medio de la aplicación del modelo PCM a las facetas calificador y tarea es posible ampliar el análisis que se realizaba en relación con las tablas 19 y 20. En la columna 7 de las mencionadas tablas se informaba de los pasos entre las puntuaciones adyacentes de los calificadores de manera independiente en las tres tareas agrupadas de las pruebas de los dos exámenes. De este modo era posible comparar la localización de los valores de paso de cada una de las bandas de calificación entre los calificadores. Con el nuevo modelo híbrido (tablas 50, 51, 52 y 53) es posible, además, comparar si la posición de los pasos entre puntuaciones es similar en cada una de las tareas del mismo calificador o si existen diferencias entre las tareas.

El estudio pormenorizado de cada uno de los calificadores haría excesivamente prolija la presentación de los resultados de las conclusiones y alargaría excesivamente el trabajo. Hemos preferido centrar el estudio en uno de los calificadores, el 4, del que habíamos señalado que asignaba en algunos casos las calificaciones 1 y 2 de forma incongruente (comenzaba a utilizar la calificación 1 a los ,49 *logit*, mientras que el valor de paso de la puntuación 2 se localizaba en –3,14 *logit*). En las tablas 50 (2 de 3) y en la 51 se constata que el valor de paso de las puntuaciones adyacentes 1–2 de este calificador en la prueba del examen D031se sitúa en dos de las tres tareas: la 2 y la 3, en una posición en el mapa de *logit* considerablemente más elevada que el valor de paso de las calificaciones 0–1 en esas mismas tareas.

En la prueba del examen D032 este calificador extrema su tendencia a no utilizar las calificaciones 0 y 1 (emplea únicamente la puntuación 1 en una ocasión) con lo que prácticamente también convierte la calificación de todas las tareas en una evaluación dicotómica. Como consecuencia de este hecho,

no se encuentran en esta prueba las contradicciones que se observaban en la anterior en el tratamiento de las calificaciones 1 y 2 en las tareas 2 y 3.

7.2.3. Análisis con el modelo PCM aplicado a las facetas calificador y atributo

Pese a disponer de bastantes datos acerca del proceder del equipo de calificadores en las tres tareas, desconocemos todavía cuál ha sido su comportamiento individual al utilizar cada uno de los atributos en el conjunto de tareas. Por medio de una nueva modificación en el programa FACETS es posible realizar este estudio; para ello es preciso combinar un modelo PCM aplicado a las facetas calificador y tarea con un modelo RSM aplicado al resto de facetas.

En las tablas 54 y 56 se localizan los valores de paso de cada una de las bandas de puntuación de los tres atributos utilizadas para la calificación de las tareas conjuntamente. De este modo es posible estudiar si la severidad o la benevolencia de los calificadores es similar en los diferentes atributos o existen diferencias significativas de actuación por parte de los examinadores entre ellas. En las tablas 55 y 57 se resumen los estadísticos más relevantes relativos al uso que cada calificador hace de los atributos en las tareas de cada una de las pruebas.

Tabla 54 (1 de 3)
Mapa de la variable EE mayo de 2012
***Logit* + Valores de paso de los calificadores**
Modelo de examen D031

```
+-----------------------------------------------------------------+
|Measr|+Candidato|-Calificador|-Tarea   |-Atributo              |
|-----+----------+------------+---------+-----------------------+
|   8 + *        +            +         +                       +
|     |          |            |         |                       |
|     | .        |            |         |                       |
|   7 + .        +            +         +                       +
|     | .        |            |         |                       |
|     | .        |            |         |                       |
|   6 + .        +            +         +                       +
|     | *        |            |         |                       |
|     | *.       |            |         |                       |
|   5 +          +            +         +                       +
|     | *.       |            |         |                       |
|     | **.      |            |         |                       |
|   4 + **.      +            +         +                       +
|     | ***.     |            |         |                       |
|     | **.      |            |         |                       |
|   3 + ***      +            +         +                       +
|     | ****     |            |         |                       |
|     | *******. | 7          |         |                       |
|   2 + ****.    +            +         +                       +
|     | ****.    |            |         |                       |
|     | *****    | 10         |         |                       |
|   1 + ****     + 12  2      +         +                       +
|     | ***      | 8          |         |                       |
|     | ***      |            | Tarea 3 | Adec-Coh   Corr-Alc   |
*   0 * ***      * 9          * Tarea 2 *                       *
|     | ****.    | 4          | Tarea 1 |                       |
|     | ***      | 1          |         | Holistica             |
|  -1 + ***.     + 3          +         +                       +
|     | *.       | 11  5      |         |                       |
|     | **.      | 6          |         |                       |
|  -2 + ***.     +            +         +                       +
|     | *.       |            |         |                       |
|     | *.       |            |         |                       |
|  -3 + .        +            +         +                       +
|     | *.       |            |         |                       |
|     | *.       |            |         |                       |
|  -4 + .        +            +         +                       +
|     | .        |            |         |                       |
|     | .        |            |         |                       |
|  -5 + .        +            +         +                       +
|-----+----------+------------+---------+-----------------------+
|Measr| * = 3    |-Calificador|-Tarea   |-Atributo              |
+-----------------------------------------------------------------+
```

Measr = Medición Rasch (*Logit*)

Tabla 54 (2 de 3)
Mapa de la variable EE mayo de 2012
***Logit* + Valores de paso de los calificadores**
Modelo de examen D031
Calificadores 1-6

Measr	1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	6.1	6.2	6.3
8	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)
7																		
6																		
5																		
4																		
3																		
2																		
1																		
0																		
-1																		
-2																		
-3																		
-4																		
-5	(1)	(1)	(0)	(1)	(1)	(0)	(1)	(1)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(1)	(1)	(0)

Measr	S.1	S.2	S.3	S.4	S.5	S.6	S.7	S.8	S.9	S.10	S.11	S.12	S.13	S.14	S.15	S.16	S.17	S.18

Measr = Medición Rasch (*Logit*)
1.1, 1.2, 1.3, 2.1 = Calificador 1, Atributo: Adec-Coh; Calificador 1, Atributo: Corr-Alc; Calificador 1, Atributo: Holistica; Calificador 2, Atributo: Adec-Coh…

Tabla 54 (3 de 3)
Mapa de la variable EE mayo de 2012
***Logit* + Valores de paso de los calificadores**
Modelo de examen D031
Calificadores 7-12

Measr	7.1	7.2	7.3	8.1	8.2	8.3	9.1	9.2	9.3	10.1	10.2	10.3	11.1	11.2	11.3	12.1	12.2	12.3
8	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)
7														---				
6													---					
5						---			---							---		---
4					---													
3		---	---								---	---			2		2	
2							---	2	2	2			2	2				
1						2									2			
* 0 *	2	2	2	2	2		2	---		---	2	2				---	2	---
-1					---								---					
-2				---		1		1	1	1				---	---	1		1
-3		---	---				---				---	---	1	1			---	
-4								---					---					
-5	(1)	(1)	(1)	(1)	(1)	(0)	(1)	(0)	(0)	(0)	(1)	(1)	(0)	(0)	(0)	(0)	(1)	(0)
Measr	7.1	7.2	7.3	8.1	8.2	8.3	9.1	9.2	9.3	10.1	10.2	10.3	11.1	11.2	11.3	12.1	12.2	12.3

Measr = Medición Rasch (*Logit*)
7.1, 7.2, 7.3, 7.1 = Calificador 7, Atributo: Adec-Coh; Calificador 7, Atributo: Corr-Alc; Calificador 7, Atributo: Holistica; Calificador 8, Atributo: Adec-Coh …

Tabla 55
Uso de las puntuaciones por los calificadores
Modelo de examen D031

	Puntuación	Frecuencia absoluta			Frecuencia Relativa			Promedio			*Outfit*			Valor de paso			SE		
		Atributo			Atributo			Atributo			Atributo			Atributo			Atributo		
		Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol
Calif. 1	0	--	--	2	--	--	2%	--	--	-3,49	--	--	,2	--	--	--	--	--	--
	1	16	16	12	13%	13%	9%	-1,24	-2,38	-1,52	1,5	,9	,2	--	--	-3,21	--	--	,66
	2	61	78	70	48%	61%	54%	,37	,77	1,46	,8	,7	,5	-2,06	-2,67	-,91	,35	,37	,33
	3	50	33	45	39%	26%	35%	3,87	4,35	5,05	,6	,4	1,0	2,06	2,67	4,12	,26	,27	,28
Calif. 2	0	--	--	2	--	--	1%	--	--	-2,77	--	--	,6	--	--	--	--	--	--
	1	20	16	19	14%	11%	13%	-2,66	-3,03	-2,11	1,2	,9	,3	--	--	-4,29	--	--	,64
	2	103	110	110	74%	79%	77%	,31	,10	1,38	,9	1,1	,9	-3,18	-3,58	-1,50	,33	,34	,29
	3	17	14	11	12%	10%	8%	3,61	3,19	4,98	,6	1,0	1,7	3,18	3,58	5,79	,31	,35	,46
Calif. 3	0	--	--	1	--	--	1%	--	--	-,36	--	--	,7	--	--	--	--	--	--
	1	22	11	12	17%	9%	9%	-,31	-1,62	-,19	2,1	,8	,5	--	--	-2,65	--	--	,79
	2	41	62	41	32%	48%	32%	,72	,81	1,08	1,6	,7	,3	-1,14	-2,15	-,09	,34	,37	,32
	3	64	55	75	50%	43%	58%	3,52	3,45	4,34	,6	1,1	1,1	1,14	2,15	2,74	,26	,24	,25
Calif. 4	0	4	4	4	3%	3%	3%	-1,04	-1,28	-,14	1,2	3,2	8,7	--	--	--	--	--	--
	1	4	1	2	3%	1%	1%	-1,06*	2,52*	-,97*	,3	,0	,5	-,59	,02	-,68	,53	,65	1,03
	2	77	115	110	57%	85%	81%	1,13	1,23	2,29	1,0	1,0	1,1	-2,11	-4,36	-3,52*	,39	,57	,76
	3	50	15	19	37%	11%	14%	2,67	3,07	4,44	1,4	1,5	,4	2,70	4,34	4,20	,23	,33	,24
Calif. 5	0	1	1	1	1%	1%	1%	-1,91	-2,15	-1,01	1,3	,4	2,0	--	--	--	--	--	--
	1	36	14	25	28%	11%	20%	-,23	-1,90	-,24	1,5	,3	1,3	-5,07	-3,99	-4,40	1,10	,69	-4,40
	2	67	79	71	52%	62%	56%	2,74	1,74	3,45	1,4	1,1	1,5	,36	-,81	-,25	,29	,32	-,25
	3	24	34	30	19%	27%	24%	4,49	4,19	5,34	1,2	2,2	,8	4,71	4,80	4,65	,28	,29	4,65
Calif. 6	0	--	--	2	--	--	3%	--	--	-2,02	--	--	,4	--	--	--	--	--	--
	1	3	2	2	4%	3%	3%	-1,04	-2,55	-1,07	,09	,3	,1	--	--	-1,24	--	--	,92
	2	28	44	33	37%	58%	42%	1,66	1,70	2,31	2,0	3,5	1,4	-2,05	-3,10	-2,00	,68	,81	,72
	3	45	30	41	59%	38%	53%	3,64	3,98	4,91	1,6	1,2	,9	2,05	3,10	3,24	,31	,31	,31
Calif. 7	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	48	23	21	36%	17%	16%	-3,21	-3,65	-3,18	2,0	,8	1,3	--	--	--	--	--	--
	2	84	98	104	62%	73%	77%	-,17	-1,25	,07	,4	1,9	,8	-3,25	-2,92	-3,41	,27	,25	,32
	3	3	14	10	2%	10%	7%	1,10	,61	1,60	,6	2,7	1,1	3,25	2,92	3,41	,44	,43	,34
Calif. 8	0	--	--	9	--	--	6%	--	--	-2,50	--	--	2,5	--	--	--	--	--	--
	1	43	7	10	32%	5%	7%	-1,69	-3,85	-2,47	2,1	,2	,3	--	--	-2,08	--	--	,41
	2	76	114	111	56%	84%	77%	,35	-,37	,83	,9	1,0	1,0	-2,15	-3,81	-2,42	,27	,35	,32
	3	16	14	14	12%	10%	10%	2,46	2,13	3,38	,4	1,8	1,1	2,15	3,81	4,50	,26	,40	,35
Calif. 9	0	--	5	2	--	4%	1%	--	-3,75	-2,40	--	,5	1,2	--	--	--	--	--	--
	1	43	40	45	32%	30%	34%	-2,32	-1,85	-1,37	1,7	,8	,5	--	-4,45	-5,00	--	,41	,73
	2	55	61	58	41%	46%	43%	1,12	,88	2,34	1,1	2,0	,4	-1,99	,03	,38	,30	,28	,29
	3	36	28	29	27%	21%	22%	4,27	4,37	5,76	,3	1,7	,5	1,99	4,42	4,62	,27	,34	,31
Calif. 10	0	3	--	--	3%	--	--	-5,40	--	--	,0	--	--	--	--	--	--	..	--
	1	23	28	20	21%	26%	19%	-2,21	-2,55	-1,78	,4	1,7	3,0	-3,76	--	--	,41	..	--
	2	66	71	79	59%	66%	73%	,05	,27	1,22	2,7	,8	1,5	-,55	-2,95	-3,23	,27	,32	,44
	3	19	9	9	17%	8%	8%	2,78	3,58	4,66	3,7	,2	,1	4,31	2,95	3,23	,45	,35	,31
Calif. 11	0	1	1	1	1%	1%	1%	-2,03	-2,27	-1,13	,6	1,0	1,1	--	--	--	--	--	--
	1	10	15	3	8%	11%	2%	-1,05	-1,47	-1,83*	1,2	1,5	,1	-3,91	-4,77	-2,24	,91	1,12	1,13
	2	96	105	89	73%	80%	67%	2,09	2,51	2,47	,9	1,0	1,3	-2,02	-2,03	-3,02	,36	,35	,58
	3	25	11	39	19%	8%	30%	5,85	5,77	6,15	1,3	1,5	1,2	5,93	6,80	5,25	,31	,37	,27
Calif. 12	0	1	--	1	1%	--	1%	-3,47	--	-2,24	,2	--	,8	--	--	--	--	--	--
	1	32	55	40	26%	44%	32%	-2,64	-1,82	-1,11	,1	6,9	,7	-5,14	--	-5,25	,57	--	,87
	2	70	66	66	70%	53%	52%	,92	1,67	1,94	2,0	1,8	1,1	,33	-3,13	,46	,26	,32	,27
	3	22	4	19	18%	4%	15%	3,63	4,00	4,50	3,5	,1	1,4	4,81	3,13	4,79	,45	,31	,36

Ad-C: Adecuación y coherencia
C-Al: Corrección y alcance
Hol: Holística

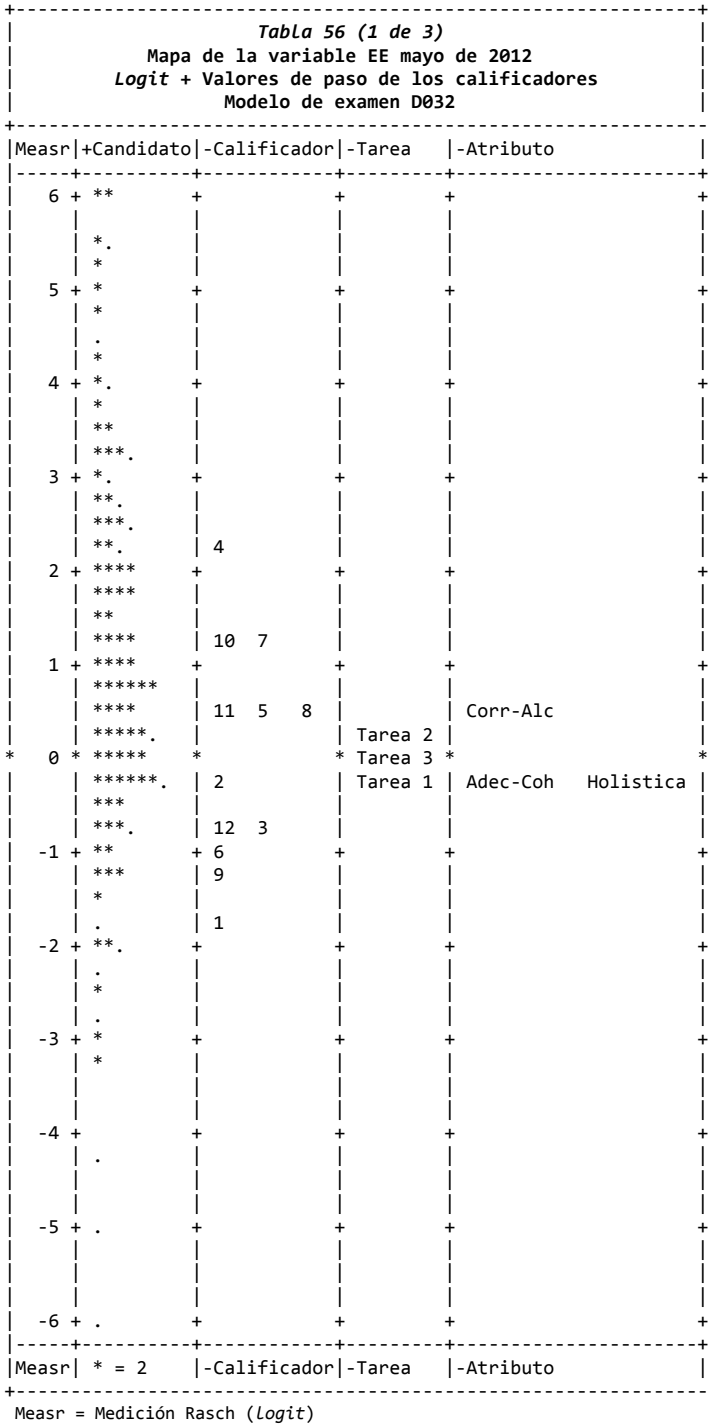


Tabla 56 (1 de 3)
Mapa de la variable EE mayo de 2012
Logit + Valores de paso de los calificadores
Modelo de examen D032
Measr
+Candidato
-Calificador
-Tarea
-Atributo
Corr-Alc
Tarea 2
Tarea 3
Tarea 1
Adec-Coh
Holistica
Measr = Medición Rasch (logit)

Tabla 56 (2 de 3)
Mapa de la variable EE mayo de 2012
***Logit* + Valores de paso de los calificadores**
Modelo de examen D032
Calificadores 1-6

Measr	1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3	4.1	5.1	5.2	5.3	6.1	6.2	6.3
6	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)
5					---	---										
4									---	---						
3		---	---	---						---			---			
2	---				2	2	---	---						---		---
1									2							
0	2	2	2	2	2	*	*	2	2	*	2	2	2	2	2	2
-1					---	---										
-2	---						---	---							---	---
					1	1					---	---				
-3		---								---			---			
-4						---										
-5																
-6	(1)	(1)	(1)	(1)	(0)	(0)	(1)	(1)	(0)	(1)	(1)	(1)	(1)	(1)	(1)	(1)

Measr	S.1	S.2	S.3	S.4	S.5	S.6	S.7	S.8	S.9	S.10	S.13	S.14	S.15	S.16	S.17	S.18

Measr = Medición Rasch (*logit*)
1.1, 1.2, 1.3, 2.1 = Calificador 1, Atributo: Adec-Coh; Calificador 1, Atributo: Corr-Alc; Calificador 1, Atributo: Holistica; Calificador 2, Atributo: Adec-Coh…

Tabla 56 (3 de 3)
Mapa de la variable EE mayo de 2012
***Logit* + Valores de paso de los calificadores**
Modelo de examen D032
Calificadores 7-12

Measr	-[···]-	7.1	7.3	8.1	8.2	8.3	9.1	9.2	9.3	10.1	10.2	10.3	11.1	11.2	12.1	12.2	12.3
6	-[···]-	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)
	-[···]-																
	-[···]-					---											
5	-[···]-									---							
	-[···]-																
	-[···]-																
4	-[···]-						---					---					
	-[···]-																
	-[···]-							---	---		---			---			
3	-[···]-												---				
	-[···]-				---											---	
	-[···]-			---		2	2								---		---
2	-[···]-	---	---														
	-[···]-									2							
	-[···]-							2	2								
	-[···]-																
1	-[···]-						---										
	-[···]-																
	-[···]-							---									
* 0	*-[···]-*	2 *	2 *	2 *	2 *	2 *	*	*	*	*	2 *	2 *	2 *	2 *	2 *	2 *	2 *
	-[···]-								---								
	-[···]-																
	-[···]-																
-1	-[···]-					---				---							
	-[···]-																
	-[···]-							1									
	-[···]-						1		1								
-2	-[···]-	---	---														
	-[···]-			---											---		---
	-[···]-				---	1				1						---	
	-[···]-																
-3	-[···]-												---				
	-[···]-							---	---		---			---			
	-[···]-																
	-[···]-									---		---					
-4	-[···]-																
	-[···]-					---											
	-[···]-						---										
-5	-[···]-																
	-[···]-																
	-[···]-																
-6	-[···]-	(1)	(1)	(1)	(1)	(0)	(0)	(0)	(0)	(0)	(1)	(1)	(1)	(1)	(1)	(1)	(1)

Measr	-[···]-	S.19	S.21	S.22	S.23	S.24	S.25	S.26	S.27	S.28	S.29	S.30	S.31	S.32	S.34	S.35	S.36

Measr = Medición Rasch (*Logit*)
7.1, 7.2, 7.3, 7.1 = Calificador 7, Atributo: Adec-Coh; Calificador 7, Atributo: Corr-Alc; Calificador 7, Atributo: Holistica; Calificador 8, Atributo: Adec-Coh…

Tabla 57
Uso de las puntuaciones por los calificadores
Modelo de examen D032

	Puntuación	Frecuencia absoluta			Frecuencia Relativa			Promedio			*Outfit*			Valor de paso			SE		
		Atributo			Atributo			Atributo			Atributo			Atributo			Atributo		
		Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol	Ad-C	C-Al	Hol
Calif. 1	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	5	5	3	6%	6%	3%	-2,83	-2,95	-3,92	,4	1,5	,1	--	--	--	--	--	--
	2	31	62	49	34%	69%	54%	1,15	1,38	1,63	,4	1,7	1,4	-1,88	-3,24	-3,07	,63	,67	,79
	3	54	23	38	60%	26%	42%	3,56	3,22	3,93	,8	,9	,8	1,88	3,24	3,07	,27	,30	,27
Calif. 2	0	--	2	2	--	3%	3%	--	-5,21	-4,40	--	,0	,3	--	--	--	--	--	--
	1	9	11	12	12%	15%	16%	-1,99	-1,57	-,80	2,1	,5	,7	--	-3,48	-4,06	--	,74	,87
	2	58	59	58	75%	79%	77%	1,09	,55	1,43	1,1	1,3	,9	-2,87	-1,35	-1,07	,60	,33	,36
	3	10	3	3	13%	4%	4%	2,77	2,77	3,10	,3	1,3	,9	2,87	4,83	5,13	,32	,77	,62
Calif. 3	0	--	--	1	--	--	1%	--	--	,56	--	--	1,2	--	--	--	--	--	--
	1	15	17	9	17%	19%	10%	--,37	-1,58	-1,31*	2,8	1,3	,3	--	--	-2,77	--	--	,67
	2	47	49	45	53%	55%	50%	,90	,41	,77	1,0	,6	,8	-1,87	-1,96	-,55	,43	,36	,32
	3	27	23	35	30%	26%	39%	3,54	3,09	3,23	,5	,5	2,2	1,87	1,96	3,32	,28	,30	,32
Calif. 4	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	1	--	--	1%	--	--	-1,24	--	--	,2	--	--	--	--	--	--	--	--
	2	57	75	74	59%	78%	77%	-,88	-1,35	-,58	,8	2,1	3,3	-3,34	--	--	,38	--	--
	3	38	21	22	40%	22%	23%	1,07	,78	1,61	7,7	,2	,1	3,34	--	--	,42	--	--
Calif. 5	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	13	16	8	-1,44	16%	8%	-1,44	-2,60	-1,98	1,4	,5	,8	--	--	--	--	--	--
	2	63	60	65	,44	60%	66%	,44	-,04	,34	,8	1,0	,7	-2,56	-2,37	-2,82	,38	-2,37	,41
	3	23	24	26	3,46	24%	26%	3,46	2,37	3,35	,5	1,3	,7	2,56	2,37	2,82	,29	2,37	,30
Calif. 6	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	2	--	1	4%	--	2%	-,10	--	-1,33	,9	--	,3	--	--	--	--	--	--
	2	21	25	21	37%	44%	37%	1,16	,40	1,08	,8	4,8	1,0	-1,75	--	-2,12	,73	--	,81
	3	34	32	35	60%	56%	61%	3,11	2,38	3,17	1,1	,4	1,3	1,75	--	2,12	,33	--	,33
Calif. 7	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	33	19	15	28%	16%	13%	-1,02	-1,92	-1,47	2,2	,5	,8	--	--	--	--	--	--
	2	72	72	79	62%	62%	68%	,43	-,82	,02	1,1	1,8	,8	-2,01	-1,93	-2,19	,28	,24	,30
	3	12	26	23	10%	22%	20%	1,01	,93	1,92	,4	2,0	,9	2,01	1,93	2,19	,28	,32	,28
Calif. 8	0	--	--	1	--	--	1%	--	--	-2,63	--	--	,2	--	--	--	--	--	--
	1	32	22	18	32%	22%	18%	-,95	-2,69	-1,48	3,0	,7	,8	--	--	-4,27	--	--	,61
	2	56	64	71	55%	63%	70%	,38	-,52	,25	3,3	,8	1,7	-2,07	-2,54	-,97	,32	,30	,27
	3	13	15	12	13%	15%	12%	3,25	3,20	3,84	,3	,6	3,4	2,07	2,54	5,24	,33	,41	,53
Calif. 9	0	1	5	5	1%	5%	5%	-2,34	-3,40	-2,68	1,1	,6	1,1	--	--	--	--	--	--
	1	22	20	16	22%	20%	16%	-,45	-,66	-,16	,7	,7	1,2	-4,61	-3,21	-3,07	,97	,59	,68
	2	38	39	42	38%	39%	42%	2,33	1,83	2,41	,9	1,5	1,3	,94	,20	-,18	,33	,33	,40
	3	38	35	36	38%	35%	36%	4,70	3,93	4,83	1,1	1,1	,6	3,67	3,02	3,25	,29	,30	,28
Calif. 10	0	2	--	--	3%	--	--	-4,25	--	--	,4	--	--	--	--	--	--	--	--
	1	11	21	13	15%	29%	18%	-3,39	-3,05	-3,07	2,0	2,0	2,2	-3,60	--	--	,60	--	--
	2	57	50	57	76%	68%	78%	,43	,07	,59	1,5	,5	,7	-1,28	-3,34	-3,67	,33	,40	,52
	3	5	2	3	7%	3%	4%	2,32	1,67	2,67	2,2	,4	,3	4,89	3,34	3,67	,73	,53	,45
Calif. 11	0	--	--	--	--	9--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	4	8	--	4%	9%	--	-1,96	-2,48	--	,5	,7	--	--	--	--	--	--	--
	2	58	67	56	64%	74%	63%	,06	-,45	-,13	,7	,9	6,4	-2,99	-3,25	--	,44	,38	--
	3	28	15	33	31%	17%	37%	2,62	2,87	2,20	2,8	1,8	,4	2,99	3,25	--	,37	,42	--
Calif. 12	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	1	3	7	3	3%	8%	3%	-1,24	-1,25	-,79	,6	1,3	,8	--	--	--	--	--	--
	2	41	59	41	46%	66%	46%	1,35	,90	1,37	,8	,7	,8	-2,25	-2,57	-2,25	,61	,51	,62
	3	51	24	46	51%	27%	51%	3,21	3,59	3,24	1,5	,7	1,4	2,25	2,57	2,25	,27	,29	,27

Ad-C: Adecuación y coherencia
C-Al: Corrección y alcance
Hol: Holística

Por medio de este nuevo análisis es posible detectar que en el calificador seleccionado como ejemplo del estudio, el 4, en la prueba del examen D031 los pasos entre las puntuaciones 1-2 se localizan en la escala de logit en una posición inferior al de los valores adyacentes 0-1 en los tres atributos. A pesar del escaso porcentaje de utilización de las puntuaciones 0 y 1 por parte de este calificador, estás disfunciones no son deseables ya que ponen de manifiesto que algunos de los candidatos calificados por este calificador han tenido que demostrar mayor competencia para que este calificador les asignen un 1 que para que les puntúe con un 2.

En la prueba del examen D032 este calificador utiliza en una única ocasión la puntuación 1 y en ninguna 0 por lo que convierte la tarea de calificación en una cuestión dicotómica y, en consecuencia, o no conocemos los valores de paso en dos de los atributos y el que conocemos no resulta significativo al haberse hallado únicamente con una única puntuación.

7.3. Análisis realizado con el modelo PCM aplicado a tres facetas: calificador, atributo y tarea

Tras los análisis realizados con el modelo RSM y con los modelos híbridos en los que se ha combinado el análisis con el modelo PCM con un modelo RSM aplicado a una o dos facetas, es posible presentar un nuevo análisis en el que se combina un modelo PCM aplicado a las tres facetas: calificador, atributo y tarea.

La información que proporciona este modelo de análisis puede resultar excesivamente prolija y detallada, ya que proporciona datos estadísticos acerca del valor de paso de cada una de las puntuaciones en cada uno de los atributos utilizadas para calificar cada una de las tareas. En consecuencia, el mapa de la variable resultante resulta excesivamente extenso para

reproducirlo. De las 36 columnas que proporciona el programa estadístico para representar los valores de paso al utilizar el último modelo utilizado anteriormente se pasan a 108 columnas. La utilización de este modelo podría resultar útil para ampliar nuevamente el foco (ampliación máxima) en el análisis de determinadas cuestiones sobre las que se quisiera completar la información. No obstante, esbozada la posibilidad, liberamos al lector de padecer la lectura de este pormenorizado modelo de análisis que resultaría excesivamente extenso y tedioso.

RESUMEN Y CONCLUSIONES

En el **primer capítulo** he analizado el nacimiento de la certificación en el ámbito de las segundas lenguas en Europa, concretamente en el Reino Unido, a mediados del siglo XIX. También he señalado que la tardía creación en el ámbito hispánico de un organismo de titularidad pública que tuviera como principal objetivo promocionar y enseñar la lengua española y difundir la cultura española e hispanoamericana en todo el mundo hizo que otras instituciones —el Ministerio de Educación, Cultura y Deporte y la Universidad de Salamanca— tuvieran que asumir la responsabilidad de cubrir esta carencia. De manera independiente y con clara anticipación de la Universidad de Salamanca, ambas instituciones comenzaron, a finales de la década de los años 80 del siglo pasado, a diseñar las primeras pruebas certificativas del nivel de competencia y dominio del idioma español. La Universidad de Salamanca terminó asumiendo el liderazgo en España y desde aquel momento hasta ahora ha colaborado, junto con otras instituciones europeas (e hispanoamericanas, el algún caso), en diversos proyectos relacionados con la evaluación de segundas lenguas: gestación de ALTE, Linguaskill, CommuniCAT, BULATS, SuveyLang, Encuesta Europea de Competencia Lingüística, SIELE… Dos momentos clave en este proceso fueron la firma de un convenio de colaboración entre el Ministerio de Educación y Ciencia y la Universidad de Salamanca en 1991, en virtud del cual los DELE son elaborados por la Universidad de Salamanca[8], y la actualización que se hizo de los mismos al vincularlos con la escala de niveles del *Marco de referencia* en 2008, ya en colaboración con el Instituto Cervantes.

También me he centrado en la cuestión terminológica relacionada con este tipo de certificaciones, tema que aún no está resuelto, y he constatado que son

[8] En el año 2014 la Universidad de Salamanca ganó la licitación pública de los Servicios de gestión para el Sistema de Certificación del Instituto Cervantes y es la institución encargada de elaborar las formas de examen y de calificar las pruebas de Expresión e interacción escritas de los DELE.

varias las posibilidades que se pueden encontrar en la literatura especializada: proficiencia, aptitud, de dominio. He considerado que la denominación que convoca mayor consenso actualmente es la de "evaluación de dominio", y ha sido esta la que hemos utilizado a lo largo de nuestro trabajo.

En el **segundo capítulo** he tratado de contextualizar la aparición del *Marco de referencia*, dada la capital importancia que este documento ha tenido y tiene en el diseño actual de los exámenes de dominio certificativos en general y de los DELE en particular. He analizado con brevedad los principales cambios que se han producido desde mediados del siglo XX en relación con la determinación del nivel lingüístico en segundas lenguas en Estados Unidos y Europa para comprender mejor la importancia de la aparición de este documento. La novedad de su enfoque —centrado en la acción—, así como las dos grandes dimensiones en las que se organizan los elementos que presenta: la horizontal, compuesta por las categorías generales o parámetros que describen el uso de la lengua y la habilidad para utilizarla, y la vertical, constituida por los niveles comunes de referencia que describen el grado de dominio lingüístico del usuario de la lengua, ha hecho que desde su aparición se haya convertido en un referente tanto en el ámbito de la docencia como en el de la evaluación de lenguas en todo el ámbito europeo. El número de trabajos que ha generado es ingente. Entre los proyectos que se desarrollaron previos a su publicación, destacamos las definiciones de requisitos mínimos de grados de dominio lingüístico que se hicieron en los estudios que se comenzaron a publicar a partir de mediados de los años 70: *Threshold Level* (Nivel umbral, en español), *Waystage* (*Plataforma*) y *Vantage* (*Ventaja*), las escalas de dominio de Eurocentres, las del Proyecto suizo, los proyectos *Can Do statements* de ALTE y DIALANG. De forma paralela a la redacción del *Marco de referencia*, y tras su publicación, surgieron proyectos como el *Porfolio Europeo de Lenguas* (PEL) y complementos al *Marco de referencia* como el *Manual*, el suplemento del *Manual*, las fichas o formularios detallados para el estudio de los capítulos 4 y 5 del *Marco de referencia* y material adicional promovido por el

Consejo de Europa relacionado con el mantenimiento de los estándares de las lenguas, contextos y administraciones por medio del juicio de expertos y las escalas TRI: las *Checklist* de ALTE, las escalas elaboradas por el Dutch CEFR Construct Group Project, el material en DVD con ejemplos de producciones orales para ilustrar los seis niveles del *Marco de referencia*, el *Plan curricular del Instituto Cervantes*, el *anual for Language Test* redactado por ALTE, etc.

En el **capítulo 3** me he centrado en los dos textos que resultan claves para el diseño y elaboración de las especificaciones de una prueba de nivel: el *Marco de referencia* y el *Plan curricular*. Pero el *Marco de referencia* no es un repertorio de normas y directrices acerca de cómo debe realizarse la evaluación de segundas lenguas; consiste en un instrumento de reflexión para los responsables del diseño de pruebas de dominio y ofrece informaciones acerca de los tipos de evaluación y los procedimientos que pueden emplearse. De ahí que haya considerado conveniente analizar cuáles son los principales aspectos que es preciso considerar para elaborar las especificaciones de una prueba de dominio. Para que el proceso de redacción se realice de forma coherente, es necesario describir y detallar todos los aspectos susceptibles de ser evaluados, así como explicitar los diversos elementos de las competencias que los candidatos deberán activar al realizar las actividades comunicativas de la lengua durante la realización de las tareas. Para ello, es preciso tener en cuenta, entre otros elementos, el modelo de competencia comunicativa utilizado, las secciones del *Marco de referencia* y del *Plan curricular* que es preciso consultar para redactar los objetivos generales y específicos, la redacción de las tareas y de las actividades de lengua, y la tipología textual utilizada.

En el **capítulo 4** he reflexionado acerca de las pruebas de desempeño (la Expresión escrita es un claro ejemplo de este tipo de evaluación) que se utilizan habitualmente en la evaluación del nivel de competencia en una lengua extranjera y en las que es primordial la participación del examinador en el proceso de calificación, especialmente cuando la prueba de ejecución

forma parte de un examen de altas consecuencias o de alto impacto, como es el caso del Diploma de Español Nivel A2, integrado en el sistema de certificación de los Diplomas de Español como Lengua Extranjera. Los test de desempeño o de ejecución presentan ventajas frente a otras formas de evaluación: son fieles, transparentes, útiles, tienen un alto grado de autenticidad, exigen mayor capacidad cognitiva, cobertura en profundidad y respuesta elaborada por el propio candidato. Su principal desventaja es que tienen un mayor coste.

Las calificaciones que reciben los candidatos en una prueba de desempeño no dependen únicamente de su nivel de competencia, sino que hay que considerar otros aspectos como el comportamiento de los calificadores, la dificultad de las tareas y la apropiada definición de los atributos que se utilizan en el proceso de calificación. Además de la correcta elaboración de las tareas de la prueba[9], es clave que los calificadores dispongan de unos criterios de calificación claros y adecuadamente desarrollados. Uno de los aspectos esenciales en el diseño de los criterios de calificación es el desarrollo de los descriptores de las escalas de calificación. Para una adecuada redacción de estas escalas hemos señalado que es preciso utilizar una combinación de métodos en su desarrollo en un proceso complementario y acumulativo: enfoque intuitivo, cualitativo y cuantitativo y seguir las siguientes fases: intuitiva, en la que se decide el sistema de redacción del borrador de escalas que se va a seleccionar; cualitativa, en la que se puede comenzar el proceso de dos modos diferentes, con descriptores o con muestras de actuación, y cuantitativa, en la que el objetivo es cuantificar el material comprobado anteriormente de manera cualitativa. En esta última fase es en la que el *Marco de referencia* propone la utilización de la TRI y, en concreto, de su modelo más directo y potente, el modelo de Rasch (Consejo de Europa 2002b: 202). Por

[9] Este aspecto solo lo hemos tratado indirectamente al hablar de la redacción de las especificaciones de la prueba, pero su estudio en profundidad excede los objetivos de este trabajo.

ese motivo, he detallado los métodos y las fases que, según el *Marco de referencia*, es necesario seguir para redactar los criterios de calificación.

Ciertamente, la influencia de la actuación de los examinadores en el proceso de calificación es determinante en la asignación de calificaciones. Las diferencias que puede haber entre los integrantes del equipo en la interpretación de las tareas, de los atributos y de las categorías de evaluación, la excesiva severidad o benevolencia de sus calificaciones y aspectos idiosincrásicos como el efecto de halo o la tendencia central contribuyen al error de la medida y pueden poner en cuestión la validez y la justicia del proceso evaluativo. Con el fin de minimizar la existencia de sesgos en el proceso de calificación es conveniente formar a los calificadores, tal y como recomienda el *Manual*. No obstante, este aspecto únicamente lo hemos esbozado en nuestro trabajo pero no hemos entrado a fondo con él. Una propuesta rigurosa y detallada de la formación que deberían seguir los examinadores de la prueba de EIE nos hubiera desviado en demasía de los objetivos previstos en nuestro trabajo.

A lo largo del **capítulo 5** he estudiado una cuestión trascendental en el trabajo: determinar qué modelo de análisis utilizar para analizar el error de medida de las calificaciones otorgadas por los examinadores de una prueba de EIE. Primeramente he considerado los procedimientos de medición basados en la Teoría Clásica de los Test, la TCT y la TG, una extensión de la anterior, dado el relevante papel que estas dos teorías han tenido en el diseño de las evaluaciones y en la cuantificación de las fuentes de error. Pero hay cuestiones que no es posible solucionar desde el enfoque clásico. Tanto la TCT como la TG no permiten obtener mediciones invariantes respecto al instrumento empleado ni determinar si las puntuaciones que obtienen los sujetos que se presentan a la prueba se deben a la severidad o benignidad de los calificadores o a la competencia de los candidatos.

Por ese motivo en el estudio he optado por utilizar un modelo psicométrico que permite obtener la separabilidad de los parámetros de los sujetos y de los ítems como es el caso del modelo dicotómico de Rasch (Andrich 1988), uno de los modelos más conocidos de la TRI. Para analizar los ítems politómicos de este tipo de pruebas en las que las respuestas habitualmente se clasifican en más de dos categorías, como es el caso de la prueba de EIE del Diploma de Español Nivel A2 he optado por utilizar las siguientes extensiones del modelo de Rasch: el Modelo de Crédito Parcial (PCM) (Wright y Masters 1982, Masters 1982 y Masters y Wright 1997), y el Modelo de Escalas de Calificación (RSM). Ambos modelos incorporan umbrales o pasos que son los valores en *logits* en los que las categorías numéricas adyacentes son equiprobables. En el modelo RSM se asume que los pasos son iguales para todos los ítems, mientras que en el modelo PCM no es necesario mantener esta suposición tan restrictiva. Finalmente, el último modelo que he descrito, es el modelo MFRM, una extensión del Modelo de Crédit Parcial (PCM) para ítems politómicos, en los que la ejecución de un sujeto es calificada por medio de un conjunto de atributos ordenados.

El modelo MFRM (Linacre 1989) es aplicable en los casos en los que existen diversas facetas de medición (candidatos, atributos, tareas y calificadores) que pueden contribuir al error de la medida y permite representar, controlando dicho error de medida, la contribución aditiva de cada faceta al *logit* o logaritmo del cociente entre la probabilidad de que una persona reciba una calificación en un atributo de una tarea y la probabilidad de que reciba la puntuación inmediatamente inferior. Este modelo de análisis me ha permitido obtener en el análisis estadísticos a nivel individual y grupal. Los estadísticos a nivel individual que hemos analizado son: medida en *logit* para cada elemento de cada faceta, error típico de medida (SE= precisión del valor estimado) e índices de ajuste entre las respuestas observadas y las predichas por el modelo. Los estadísticos a nivel grupal son: ajuste promedio, media, variabilidad y fiabilidad de las medidas de las personas, las tareas y los calificadores.

Una vez seleccionados los modelos de análisis y tras detallar la procedencia de los candidatos que se presentaron a las pruebas, en el **capítulo 6** informamos de las características de las pruebas que calificaron los examinadores: dos pruebas de EIE del Diploma de Español Nivel A2 de la convocatoria de mayo de 2012. En su calificación participaron doce examinadores (todos calificaron tareas de las dos pruebas). Tal y como indicaba al comienzo de estas conclusiones, resultó crucial la manera de distribuir las pruebas entre los calificadores. El principal objetivo que perseguía era garantizar la conectividad entre ellos, pero de forma que el proceso de calificación fuera viable; por ese motivo, deseché la posibilidad de que todos los examinadores calificaran las pruebas de todos los candidatos, por el excesivo tiempo que llevaría el proceso y por el coste económico que supondría; factores que lo hacían desaconsejable. Seguí, entonces, el procedimiento propuesto por Linacre y Wright (2002) para disminuir el número de calificaciones totales manteniendo la conexión entre candidatos, calificadores y pruebas. Según el plan propuesto, cada candidato comparte calificador con otro candidato y, al menos, dos calificadores evalúan cada prueba. En el trabajo he utilizado un diseño simple de rotación de los examinados y los calificadores.

Los dos exámenes se realizaron en la convocatoria de mayo de 2012, en dos días consecutivos: el D031 tuvo lugar el día 25 de mayo y el D032, el día 26. El diseño de rotación entre candidatos y calificadores se diseñó para 443 candidatos (259 de los que se presentaron al examen del día 25 y 184 de los del día 26). Los candidatos redactaron tres textos que fueron evaluados de forma independiente por dos calificadores, que analizaron las tres tareas de cada candidato y puntuaron en cada una de ellas tres atributos organizados en dos escalas, una analítica (Adecuación al género discursivo y coherencia y Corrección y alcance) y una holística: calificación Holística. Participaron doce examinadores en el proceso de calificación y los calificadores fueron los mismos en ambas pruebas.

El objetivo general perseguido en el **capítulo 7** fue el de ilustrar las potencialidades del modelo MFRM para obtener medidas objetivas de las facetas implicadas en la evaluación de una prueba de desempeño o ejecución: calificadores, tareas, atributos y candidatos con el propósito de visualizar en una única tabla los elementos de las diferentes facetas analizadas calibrados en la misma escala de intervalos (*logit*) con el fin de que fuera posible comparar e interpretar los resultados de la competencia de los candidatos, la severidad de los calificadores, la dificultad de las tareas y de los atributos, así como de localizar los valores de paso de los valores adyacentes en un mismo marco de referencia.

Primeramente he utilizado un modelo en el que se aplica a las cuatro facetas el modelo RSM y he analizado el mapa de la variable y el comportamiento en el proceso de calificación de los calificadores, tareas, candidatos y atributos. En relación con los errores que pueden cometer los examinadores en el proceso de calificación, he comenzado estudiando la variabilidad en severidad entre los calificadores, que ha resultado ser moderadamente alta. Además, el elevado índice RSR en las dos pruebas ha revelado que las diferencias en severidad observadas entre los calificadores eran muy fiables y el hecho de que los errores estándar fueran bastante bajos y muy uniformes entre los calificadores indicaban que la precisión de la medición en ambas pruebas era alta. Además, se detectó que la correlación entre los valores de severidad de los calificadores de ambas pruebas, en términos tanto de direccionalidad como de intensidad o de fuerza, era moderada con tendencia a fuerte. Esta variabilidad contribuye al error de la medida y decrementa la validez y la justicia del proceso evaluativo.

He constatado que resulta complicado atribuir estos elevados valores de severidad a factores concretos, aunque entre las causas que pueden influir en que un examinador valore con mayor o menor severidad la actuación de un candidato pueden ser: la cantidad de tareas que tenga que calificar y el tiempo

de que disponga para hacerlo; factores idiosincrásicos como la personalidad del calificador, y su actitud ante el proceso de calificación en el que va a participar, su experiencia… Parece que, con frecuencia, los calificadores más experimentados suelen ser más severos que los neófitos (Eckes 2011) aunque, según mi experiencia, son habitualmente los nuevos calificadores los que se muestran excesivamente severos o benévolos, mientras que los veteranos, que tienen más experiencia y han realizado cursos de formación, presentan unos valores de severidad no tan extremos. Sin embargo, he comprobado que no existen estudios suficientes acerca de los motivos que pueden ocasionar los excesos de severidad o de benignidad en los equipos de calificadores.

También he considerado que la formación de los calificadores es un aspecto esencial en el proceso de calificación, ya que mediante ella se persigue minimizar la existencia de determinados sesgos en la calificación, como el que se acaba de señalar. El *Manual* detalla cuáles son los pasos que deben seguirse para que el proceso de calificación individual se realice con las suficientes garantías. No obstante, es conveniente recordar que, tal y como Eckes (2011) indica, en muchas ocasiones la formación de los calificadores no resulta eficaz para reducir las diferencias de severidad entre ellos.

Para estudiar la fiabilidad de los calificadores he analizado la consistencia interna intra-calificador y la concordancia entre calificadores. En relación con el primero de los aspectos, y tras estudiar los estadísticos de ajuste (*Infit* y *Outfit*), he comprobado que el equipo de calificadores se ha ajustado de manera bastante aceptable al modelo y ha mostrado una alta consistencia intra-calificador en sus evaluaciones. En el estudio he observado que los valores se situaban en un rango aceptable: *Infit* entre ,72 y 1,28 y *Outfit* entre ,65 y 1,76 en la prueba del examen D031 e *Infit* entre ,66 y 1,23 y *Outfit* entre ,62 y 1,27 en la prueba del examen D032. Únicamente ha resultado destacable el moderado grado de *misfit* que presenta el calificador 6 en la prueba del examen D031 y la leve tendencia hacia *overfit* de los calificadores 1 y 2 en la

misma prueba y del 4, el 10 y de nuevo el 2, en la del D032. El segundo de los procedimientos utilizados para analizar la fiabilidad de los calificadores es la concordancia entre calificadores (Rc-rc). La correlación entre cada calificador con el resto de calificadores cuantifica el grado en el que las calificaciones de los examinadores son consistentes con las del resto. Las correlaciones son en todos los casos superiores a ,50 (el valor mínimo es el del calificador 6 en le prueba del examen D032 con un valor de ,56) y se considera que los valores superiores a 0,30 indican que la evaluación es consistente y que la ordenación que hacen de los candidatos mediante las calificaciones otorgadas es similar a la del resto de examinadores.

Por medio del error estándar he analizado los intervalos de confianza entre los que se espera que el valor de severidad se encuentre en el 95% de las ocasiones (±2SE) y he constatado que en las dos pruebas es posible establecer ocho grupos de calificadores con distinto grado de severidad.

Además de la severidad de los calificadores, también he analizado otros posibles errores de examinadores: efecto de tendencia central y efecto de halo.

La tendencia central consiste en un tipo especial de restricción de rango (Myford y Wolfe 2004a) y se produce cuando un examinador utiliza con excesiva frecuencia la banda de calificación intermedia de cada uno de los atributos y asigna en escasas ocasiones las puntuaciones altas y bajas de la escala. Este tipo de actuación por parte de un calificador ocasiona que, en los resultados de sus valoraciones, se sobreestime el nivel de competencia de candidatos con bajo nivel de competencia y se minusvalore el de los candidatos con alto nivel. Para analizar la existencia de esta tendencia resulta interesante analizar la frecuencia con la que el equipo de calificadores utiliza cada una de las bandas de calificación. En el estudio, no he detectado fuertes indicios de tendencia central en el grupo de calificadores ya que, para que se pueda afirmar que existe esta tendencia, es necesario (Myford y Wolfe 2004a) que haya: una baja frecuencia en la utilización de las categorías extremas, y

este hecho no se observa en el equipo de calificadores, ya que los valores de utilización de las categorías extremas 1 y 3 se aproximan al 40% en las dos pruebas; que el índice de fiabilidad RSR sea bajo, y en las dos pruebas está cerca de 1, y que los valores de *Infit* y *Outfit* sean extremos , mientras que los promedios en ambas prueba apenas difieren de la unidad. El análisis del valor de chi-cuadrado en ambas pruebas corrobora esta conclusión, ya que su estudio no revela la presencia de tendencia central.

En el estudio particular de los calificadores, he observado que los que utilizaron las categorías extremas en un porcentaje inferior al 25% fueron el 2 (13%) y el 4 (23%) en la prueba del examen D031 y el 2 (21%) y el 10 (25%) en la del D032. De acuerdo con estos indicadores parece que el calificador 2 presenta un efecto de tendencia central. En relación con los valores *Outfit* de las calificaciones intermedias de los examinadores de manera individual, en la prueba del examen D031 el calificador que ha presentado un *Outfit* más elevado en la puntuación 2 es el 6 (1,9), mientras que el 1 es el que lo ha tenido más bajo (,6). Wright y Linacre (1994) consideran los márgenes aceptables van de 0,6 a 1,4. De todos modos, como también advierten los mencionados autores, estos límites deben interpretarse como sugerencias y no pueden considerarse como fronteras rígidas; de manera que con estos datos no es posible confirmar la existencia de efecto de tendencia central con el estudio de este estadístico.

Otra de las fuentes de información que puede resultar útil para detectar el efecto de tendencia central a nivel individual es el análisis de las diferencias entre los pasos de las curvas características de las categorías (Myford y Wolf 2004b). Estos umbrales representan el punto en el que la probabilidad de que un candidato esté clasificado en una de las dos puntuaciones adyacentes es de un 50%. Cuando un calificador presenta efecto de tendencia central, la media de las diferencias entre los pasos de las categorías es mayor en los calificadores que puntúan con un efecto de tendencia central (Eckes 2011), es

decir, los valores de paso se encuentran muy dispersos. Además, habitualmente, los calificadores que presentan efecto de tendencia central no suelen utilizar calificaciones extremas. En relación con el equipo de calificadores, se observa que, en líneas generales, su comportamiento es bastante homogéneo y no se detectan diferencias importantes en los valores de paso entre las tareas de cada una de las pruebas. Sin embargo, en el análisis individual, se constata que la media de los valores de paso, en la prueba del examen D031, del calificador 2 ha sido de 5,48 *logit* y que, en la prueba del examen D032, la media entre los valores de paso del calificador 10 ha sido de 6,74 *logit*. Esta dispersión puede evidenciar la presencia de un efecto de tendencia central en estos dos examinadores.

El tercero de los errores objeto del análisis realizado, el efecto de halo, se produce cuando un calificador elige con excesiva frecuencia la misma categoría en los diferentes atributos, sin tener en cuenta el diferente nivel de competencia que puede haber demostrado tener el candidato en cada uno de ellos, ni las diferencias existentes. Cuando la mayor parte de los examinadores presentan efecto de halo, las medidas de los atributos evaluados apenas varían: la diferencia entre los promedios de los atributos en la prueba del examen D031 ha sido de 0,04 puntos y entre los de la del D032, de 0,1 puntos.

La fiabilidad del índice de separación entre los atributos informa acerca de si los atributos se diferencian entre sí y si los calificadores han sido capaces de distinguirlos. Una baja fiabilidad del índice podría sugerir efecto de halo en las calificaciones (Myford y Wolf 2004b). Los valores de fiabilidad han sido de ,75 en la prueba del examen D031 y de ,91 en la del D032, lo que podría indicar un cierto efecto de halo en la prueba del examen D031.

Tras el pormenorizado estudio en relación con los calificadores, he analizado el comportamiento de las tareas y he observado que su comportamiento fue similar en ambos exámenes. Además, el error típico de la medida ha sido muy

bajo en ambas pruebas, lo que indicaba que la precisión de las estimaciones de la dificultad era muy alta. Asimismo, el hecho de que el ESR sea elevado en ambas pruebas revelaba que las tareas tenían diferente nivel de dificultad y que podía intentar garantizar un muestreo adecuado de los distintos niveles de dificultad del constructo que se evaluaban. Ninguna de las tareas se desajustaba severamente con las predicciones del modelo.

En relación con los candidatos, he observado que estos han obtenido un elevado nivel de rendimiento en ambas pruebas (algo superior a 2 *logits* en promedio) y una alta variabilidad (las desviaciones típicas de las puntuaciones también son altas). Las puntuaciones otorgadas a los candidatos han tenido también una fiabilidad muy alta (PSR superior a ,9 en las dos pruebas) y el número de candidatos que presentaban desajuste severo con las predicciones del modelo ha sido bajo (inferior en las dos pruebas al 9%).

Arriba señalaba que no existen grandes diferencias de dificultad entre los atributos utilizados para calificar las tareas y que la precisión de las estimaciones de la dificultad de los atributos es muy alta, ya que el error típico de la medida (SE) es muy pequeño y los estadísticos de fiabilidad global (ISR) son elevados. Las elevadas correlaciones atributo–escala han revelado que existía un patrón de competencia semejante en las variables evaluadas, por lo que ha sido adecuado combinarlas en una única puntuación para reflejar el rendimiento de los candidatos. La realización de un análisis con el modelo PCM aplicado a la faceta atributo ha permitido visualizar los valores de paso del conjunto de calificadores en cada uno de ellos y me permitió observar que, en líneas generales, el comportamiento de los calificadores en la utilización de las bandas de calificación de los tres atributos ha sido bastante homogéneo, aunque sí se podrían detectar ciertas tendencias.

Con el fin de obtener información complementaria a la que ya había obtenido por medio de los análisis anteriormente realizados, he aplicado una nueva modificación del modelo de análisis del programa FACETS con el fin de

combinar un modelo PCM aplicado a dos facetas con un modelo RSM aplicado al resto.

El análisis con un modelo PCM aplicado a las facetas atributo y tarea permitió conocer si los valores de paso de los atributos en las tareas eran homogéneos y constatar que el valor de paso de las puntuaciones 1–2 se ha localizado en una posición más elevada en el mapa de la variable en el atributo Adecuación y coherencia en todas las tareas de las dos pruebas, excepto en la 2 de la prueba del examen D032, en la que el paso entre estas dos calificaciones se ha localizado en una posición de alrededor de 2 *logit*, inferior a los de los otros dos atributos. Este hecho podría indicar algún problema concreto en el diseño o en la construcción de dicha tarea 2.

Al modificar las facetas analizadas por medio del modelo PCM (calificador y tarea) he podido conocer si cada uno de los examinadores había calificado de igual modo cada una de las tareas o si su proceder había variado dependiendo de la tarea que estuviera puntuando. De este modo ha sido posible ampliar el foco para conocer acerca de la tendencia idiosincrásica de los calificadores a la hora de aplicar las categorías y analizar la severidad de los examinadores y los valores de paso de las puntuaciones en cada una de las tareas para cada uno de los calificadores.

En la introducción de esta tesis doctoral me planteaba el objetivo de llegar a conocer el comportamiento de los examinadores en los procesos de calificación de pruebas de EIE. Creo, sinceramente, que las conclusiones a las que se ha llegado en este trabajo son muy esperanzadoras. Desde el campo de la Filología es fácil ver los problemas, ser conscientes de que hay calificadores que son muy rigurosos a la hora de calificar los textos escritos por los estudiantes y de que hay otros que pecan de lenidad. Pero para ser capaces de resolver el problema es necesario comprenderlo primero. Y nuestro acercamiento a la Psicometría nos ha ayudado en este propósito. El descubrimiento del modelo MFRM me ha permitido observar a las diferentes

facetas que intervienen en un proceso de califiación en el mapa de la variable, calibrar la severidad de los calificadores con independencia de los candidatos que hayan realizado la prueba, estudiar la presencia de otros tipos de errores en la calificación como la tendencia central o el efecto de halo, etc.

Su utilización es muy recomendable cuando la evaluación está mediada por calificadores: pruebas de selectividad, test universitarios, exámenes de dominio, pruebas de selección en empresas… Los procesos de medición ganan en validez, fiabilidad y transparencia y se posibilita la adecuada formación de los calificadores.

A N E J O

Escalas de calificación de la prueba de EIE del Diploma de español nivel A2

DELE A2. Expresión e interacción escritas - Escala analítica	
	Adecuación al género discursivo y coherencia
3	Escribe textos descriptivos o narrativos ordenados mediante una secuencia lineal de elementos sencillos, utilizando organizadores de la información *(primero, luego, después)* y los conectores básicos más frecuentes *(y, también, por eso, entonces, pero...)*. Utiliza de forma adecuada los signos de puntuación. Respeta las convenciones de género básicas (inicio y cierre del texto...) en cartas, mensajes, notas..., y utiliza las fórmulas básicas de cortesía (saludo, despedida...). El resultado final es un texto sencillo y cohesionado.
2	Realiza descripciones y narraciones muy breves y básicas, sobre su entorno más inmediato o aspectos de la vida cotidiana (lugares, personas, entidades y objetos), enlazando palabras con conectores muy sencillos y básicos *(y, pero, porque)*. Redacta cartas, mensajes y notas sencillas y breves sobre áreas de necesidad inmediata o para transmitir información personal muy básica, utilizando los exponentes funcionales más frecuentes, normas de cortesía elementales o fórmulas de saludo y tratamiento *(muchas gracias; hola, ¿cómo estás?)*, aunque con vacilaciones. Puede faltarle algún detalle importante. En los textos más largos pueden producirse errores: uso indebido de elementos de referencia *(llevar/traer)*, elección indebida de deícticos, vacilación en el uso de los signos de puntuación y falta de organización del texto, que hacen necesaria una relectura para su comprensión, aunque logra su objetivo.
1	El texto producido se limita a una serie de frases sencillas aisladas sobre sí mismo y sobre otras personas o temas de su entorno más próximo. En algunos casos, la información aparece desordenada o incompleta, lo que obliga a una relectura para su comprensión. Utiliza fórmulas más sencillas y cotidianas relativas a saludos, despedidas, presentaciones y expresiones del tipo *por favor, gracias, lo siento*..., que adapta al discurso con errores importantes de formulación. Hay errores de registro y faltan detalles importantes.
0	El texto no se corresponde en registro, estilo o estructura a la situación planteada ni sigue las pautas dadas. El grado de formalidad, el tono, el nivel de detalle o el léxico no son adecuados para el contexto. Los textos no mantienen una estructura organizada que permita seguir los razonamientos del candidato. En textos creativos, el estilo es inapropiado y no se siguen las indicaciones dadas. En cartas e informes, no se incluye la información necesaria o no hay un orden claro.

DELE A2. Expresión e interacción escritas - Escala analítica	
	Corrección y alcance
3	Muestra un control razonable de elementos lingüísticos básicos y estructuras habituales: distinción *ser* y *estar* (usos básicos), oraciones interrogativas y exclamativas, imperativo afirmativo y negativo, uso de las perífrasis verbales más frecuentes..., que utiliza para satisfacer necesidades inmediatas y en temas predecibles o de interés personal. Puede producirse algún error en la ortografía de las palabras (tildes en el vocabulario frecuente, abreviaturas...) pero que no interfiere en la transmisión de la idea principal del texto. Domina el vocabulario básico suficiente para desenvolverse con bastante precisión en intercambios cotidianos próximos a su entorno más inmediato. Esto le permite solicitar información, hacer valoraciones, expresar deseos, dar instrucciones. Puede cometer errores si utiliza estructuras o vocabulario más complejo.
2	Utiliza correctamente estructuras gramaticales sencillas pero sistemáticamente comete errores básicos y predecibles en este nivel: vacilación en el uso de *ser/estar/haber*, errores en las formas de los tiempos verbales regulares e irregulares, confusiones con los pronombres y adjetivos indefinidos, en las concordancias de sujeto-verbo o nombre-adyacentes... Sin embargo, suele entenderse lo que quiere transmitir siempre que el mensaje esté relacionado con una situación comunicativa frecuente. Tiene un repertorio limitado de exponentes lingüísticos, que utiliza para transmitir información básica en situaciones que respondan a necesidades muy concretas y cotidianas: pedir objetos, dar información personal, decir que le duele algo... Presenta ciertas imprecisiones que no afectan a la comunicación.
1	Muestra un control limitado de algunas estructuras gramaticales muy básicas y sencillas *(estar+gerundio, gustar+infinitivo...)* o frases cortas construidas básicamente en presente de indicativo, previamente memorizadas, relativas a necesidades básicas e inmediatas. Comete abundantes errores gramaticales y ortográficos (concordancias, error en la elección de la persona del verbo), que dificultan la comprensión del mensaje y que hacen que el texto sea incomprensible. Utiliza un repertorio de palabras muy básicas y frases aisladas que no es suficiente para transmitir la información requerida ni para que se produzca la comunicación. Constantes imprecisiones léxicas e interferencias de otras lenguas
0	Utiliza frases aisladas y palabras sin sentido. El texto producido presenta numerosos errores que hacen imposible su interpretación.

DELE A2. Expresión e interacción escritas - Escala holística	
3	Aporta a la información requerida algunos detalles que hacen que la organización y formulación del mensaje cumplan sobradamente con los objetivos comunicativos planteados. Se ajusta a la extensión y desarrolla con eficacia los puntos dados transmitiendo el mensaje con cierta precisión. Utiliza un repertorio lingüístico básico, suficiente para expresarse sobre situaciones cotidianas y temas de su interés. A pesar de algunos errores, construye oraciones sencillas y breves, con palabras clave en un discurso comprensible y claro.
2	Aporta la información requerida de forma comprensible y logra transmitir los mensajes. Se expresa de manera sencilla en intercambios de información en temas conocidos y cotidianos. Utiliza un repertorio lingüístico limitado compuesto por estructuras sintácticas y expresiones memorizadas en un discurso con errores elementales de concordancia, morfemas, etc., que dificultan la comprensión pero que dejan clara la idea general. Sigue la gran mayoría de los puntos de orientación dados, aunque alguno no esté lo suficientemente desarrollado.
1	Aporta solo algunos datos que resultan insuficientes para transmitir los mensajes. El escrito se limita a una serie de oraciones muy breves, simples, en un discurso desorganizado y con errores abundantes que dificultan la comprensión del mensaje.
0	El texto es comprensible, pero no proporciona la información requerida y resulta demasiado breve y confuso, para lograr el objetivo comunicativo planteado. Las limitaciones lingüísticas (errores e imprecisiones gramaticales, léxicas, sintácticas, ortográficas o de puntuación) provocan dificultades en la formulación de lo que quiere decir. El texto no sigue los puntos de orientación dados. Extensión por tareas: Tarea 1: El candidato escribe menos de 20 palabras. Tareas 2 y 3: El candidato escribe menos de 55 palabras.

Instituto Cervantes (2013). En: <http://diplomas.cervantes.es/informacion/guias/materiales/a2/guia_examen_dele_a2.pdf.

R E F E R E N C I A S B I B L I O G R A F Í C A S

Abad, Francisco J.; Olea, Julio; Ponsoda, Vicente y García, Carmen (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

Alderson, J. Charles (1991). "Bands and scores", en: Alderson, J. Charles y North, Brian (eds.). *Language Testing in the 1990s. Developments in ELT*. London: British Council/Macmillan, 71-86.

Alderson, J. Charles (coord.); Figueras, Neus; Kuiper, Henk; Nold, Günther; Takala, Sauli y Tardieu, Claire (2006). "Analysing tests of reading and listening in relation to the Common European Framework of Reference: the experience of the Dutch CEFR Construct Project". *Language Assessment Quarterly* 3 (1), 3-30.

Alderson, J. Charles; Clapham, Caroline y Wall, Dianne (1998). *Exámenes de idiomas*, Cambridge: Cambridge University Press. Primera edición en ingles: (1995). *Languaje Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J. Charles; Krahnke, Karl J. y Stansfield, Charles W. (eds.) (1987). *Reviews of English language Proficiency Tests*. Washington, DC: Teachers of English to Speakers of Other Languages.

ALTE (1999). *Multilingual Glossary of Language Testing Terms: Studies in Language Testing 6*. Cambridge: Cambridge University Press.

ALTE (2002). *The ALTE Can Do Project. English version. Articles and Can Do Statements Produced by the Members of ALTE 1992-2002*. ALTE. En: <http://www.cambridgeenglish.org/images/28906-alte-can-do-document. pdf> (30 de noviembre de 2015).

ALTE (2002b): *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Language examining and test development*. Estrasburgo: Consejo de Europa. En: <http://www.atriumlinguarum.org/ contenido/Guia%20Test.pdf> (30 de noviembre de 2015).

ALTE (2005a). *The CEFR Grid for Speaking, Developed by ALTE Members (input)*. v. 1.0. En: <https://www.coe.int/t/dg4/linguistic/ Source/ALTE% 20CEFR%20Speaking%20Grid%20INput51.pdf> (30 de noviembre de 2015).

ALTE (2005b). *The CEFR Grid for Speaking, Developed by ALTE Members: Report*. v. 2.0. En: <https://www.coe.int/t/dg4/linguistic/

Source/ALTE% 20CEFR%20Speaking%20Grid%20OUTput51.pdf> (30 de noviembre de 2015).

ALTE (2007a). *The CEFR Grid for Writing Task (Analysis)*. v. 3.1. En: <http://www.coe.int/t/dg4/linguistic/Source/CEFRWritingGridv3_1_analysis.doc> (30 de noviembre de 2015).

ALTE (2007b). *The CEFR Grid for Writing Task (presentation)*. v. 3.1. En: <http://www.coe.int/t/dg4/linguistic/Source/CEFRWritingGridv3_1_analysis.doc> (30 de noviembre de 2015).

ALTE (2009a). *Development and Descriptive Checklist for Tasks and Examinations. General*. ALTE. En: <http://www.alte.org/attachments/files/general_check.pdf>(30 de noviembre de 2015).

ALTE (2009b). *Individual Component Checklist. Reading*. En: <http://www.alte.org/attachments/files/reading_check.pdf> (30 de noviembre de 2015).

ALTE (2009c). *Individual Component Checklist. Writing*. En: <http://www.alte.org/attachments/files/writing_check.pdf> (30 de noviembre de 2015).

ALTE (2009d). *Individual Component Checklist. Listening*. En: <http://www.alte.org/attachments/files/listening_check.pdf> (30 de noviembre de 2015).

ALTE (2009e). *Individual Component Checklist. Speaking*. En: <http://www.alte.org/attachments/files/speaking_check.pdf> (30 de noviembre de 2015).

ALTE (2009f). *Individual Component Checklist. Structural Competence*. En: <http://www.alte.org/attachments/files/ structural_comp.pdf> (30 de noviembre de 2015).

ALTE (2009g). *Individual Component Checklist for Use with ONE Task. Reading*. En: <http://www.alte.org/attachments/files/ reading_check_onetask.pdf> (30 de noviembre de 2015).

ALTE (2009h). *Individual Component Checklist for Use with ONE Task. Structural Competence*. En: <http://www.alte.org/attachments/files/structural_comp_onetask.pdf> (30 de noviembre de 2015).

ALTE (2009i). *Individual Component Checklist for Use with ONE Task. Writing*. En: <http://www.alte.org/attachments/files/writing_check_onetask.pdf> (30 de noviembre de 2015).

ALTE (2009j). *Individual Component Checklist for Use with ONE Task. Listening*. En: <http://www.alte.org/attachments/files/listening_check_onetask.pdf> (30 de noviembre de 2015).

ALTE (2009k). *Individual Component Checklist for Use with ONE Task. Speaking*. En: <http://www.alte.org/attachments/files/speaking_check_onetask.pdf> (30 de noviembre de 2015).

ALTE (2011). *Manual for Language Test Development and Examining. For Use with the CEFR*. Estrasburgo: Consejo de Europa. En: <http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf> (30 de noviembre de 2015).

American Educational Research association; American Psychological Association y National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Andrich, David (1978). "Application of a psychometric model to ordered categories which are scored with successive integers". *Applied Psychological Measurement,* 2 (4), 581-594. En: <http://hbanaszak.mjr.uw.edu.pl/TempTxt/Andrich_1978_Application%20of%20a%20Psychometric%20Rating%20Model%20to%20Ordered%20Categories%20Which%20Are%20Scored%20with%20Successive%20Integers.pdf> (30 de noviembre de 2015).

Andrich, David (1988). *Rasch Models for Measurement*. Newbury Park, CA: Sage.

Andrich, David (1998). "Threshold, steps and rating scale conceptualization". *Rasch Measurement Transactions,* 12, 648-649.

Andrich, David (2005). "Rasch models for ordered response categories", en: Everitt, Brian Sidney y Howell, David C. (eds.). *Encyclopedia of Statistics in Behavioral Science*. Vol. 4. Nueva York, NY: Wiley, 1698-1707

Arter, Judith y McTighe, Jay (2001). *Scoring Rubrics in the Classroom: Using Performance Criteria for Assessing and Improving*. Thousand Oaks, CA: Corwin.

Bachman, Lyle F. (1990). *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

Bachman, Lyle F. y Palmer, Adrian S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Baker, Frank B. (1989). "Computer technology in test construction and processing", en: Linn, Robert L. (ed.), *Educational Measurement*. Nueva York, NY: Macmillan, 409-428.

Barbero, Mª Isabel (1999). "Gestión informatizada de bancos de ítems", en: Olea, Julio; Ponsoda, Vicente y Prieto, Gerardo (eds.). *Tests informatizados. Fundamentos y aplicaciones*. Madrid: Pirámide, 63- 83.

Baron, Joan Boykoff (1991). "Strategies for the development of effective performance exercises". *Applied Measurement in Education*, 4 (4), 305-318.

Barret, Seven (2005). "Raters and examinations", en: Alagumalai, Sivakumar; Curtis, David C. y Hungi, Njora (eds). *Applied Rasch Measurement: A Book of Exemplars – Papers in Honour of John P. Deeves*. Dordrecht: Springer, 159-177.

Birnbaum, Allan (1968). "Some latent trait models and their use in inferring an examinee's ability", en Lord, Frederic M. y Novick, Melvin R. (eds.). *Statistical theories of mental test scores*. Reading, MA: Addison–Wesley.

Bock, R. Darrell (1997). "A brief history of item response theory". *Educational Measurement: Issues and Practice*, 16 (4), 21-33.

Bond, Trevor G. y Fox, Christine M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2ª ed.). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Boomsma, Anne; Van Duijn, Marijtje A. J. y Snijders, Tom A. B. (2001). *Essays on item response theory*. Nueva York, NY: Springer,

Braun, Henry I. (1988). "Understanding scoring reliability: Experiments in calibrating essay readers". *Journal of Educational Statistics*, 13 (1), 1–18.

Breton, Gilles; Lepage, Sylvie y North, Brian (2008). *Cross-language Benchmarking Seminar to Calibrate Examples of Spoken Production in English, French, German, Italian and Spanish with Regard to the Six Levels of the* Common European Framework of Reference for Languages (CEFR). Estrasburgo: Consejo de Europa. En: <http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key_reference/Sevres_Report_2008_EN.pdf> (30 de noviembre de 2015).

Burt, Cyril (1936). "The analysis of examination marks", en: Hartog, Philip Joseph y Rhodes, Edmund Cecil (eds.). *The Marks of Examiners*. Londres: Macmillan, 245-314.

Campbell, Donald T. (1957). "Factors relevant to the validity of experiments in social settings". *Psychological Bulletin*, 54, 297-312.

Campbell, Donald T., y Stanley, Julian C. (1963). "Experimental and auasi-experimental designs for research on teaching", en: Nathaniel Lees Gage

(ed.), *Handbook of research on teaching*. Chicago, IL: Rand McNally, 171-246.

Canale, Michael (1983), "From communicative competence to communicative language pedagogy", en: Richards, Jack C. y Schmidt, Richard W. (eds.), *Language and communication*. London, Longman, 2-27. Existe traducción al español del artículo con el título: "De la competencia comunicativa a la pedagogía comunicativa del lenguaje", en: Llobera, Miquel (coord.) (1995). *Competencia comunicativa : documentos básicos en la enseñanza de lenguas extranjeras*. Madrid: Edelsa, 63-81.

Canale, Michael y Swain, Merrill (1980). "Theoretical bases of communicative approaches to second language teaching and testing", *Applied Linguistics*, 1 (1), 1-47.

Canale, Michael y Swain, Merrill (1981). "A theoretical framework for communicative competence", en: Palmer, A. S., Groot, P. G. y Trosper, S.A. (eds.). *The construct validation of tests of communicative competence*. Washington, DC: TESOL, 31-36.

Casanova, Dominique y Demeuse, Marc (2011). "Analyse des différentes facettes influant sur la fidélité d'une l'épreuve d'expression écrite d'un test de français langue étrangère". *Mesure et évaluation en éducation* 34 (1), 25-53.

Cascio, Wayne F. (1982). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, 13, 1-18.

Chalhoub-Deville, Micheline (1995). "Deriving oral assessment scales across different tests and rater groups". *Language Testing*, 12 (1), 16-33.

Chalhoub-Deville, Micheline, ed. (1999). *Issues in Computer-adaptive Testing of Reading Proficiency*. Cambridge, UK.: University of Cambridge Local Examinations Syndicate y Cambridge University Press.

Clauser, Brian E. (2000). "Recurrent issues and recent advances in scoring performance assessments". *Applied Psychological Measurement*, 24 (4), 310-324.

Committee for Out-of-school Education and Cultural Development (1971). *Linguistic Content, Means of Evaluation and their Interaction in the Teaching and Learning of Modern Languages in Adult Education: Report of a Symposium Organised at Rüschlikon, Switzerland 3 - 7 May, 1971*. Estraburgo: Consejo de Europa.

Consejo de Europa (1982). "Recommendation nº R(82) 18 of the Committee of Ministers to member States concerning modern languages", en: Anejo A de Girard, Denis y Trim, John Leslie Melville (eds.) (1988). *Project nº 12*.

'Learning and Teaching Modern Languages for Communication': Final Report of the Project Group (activities 1982-87). Estrasburgo: Consejo de Europa.

Consejo de Europa (1992), *Transparency and Coherence in Language Learning in Europe: Objectives, Evaluation, Certification,* Estrasburgo, Consejo de Europa. En: <https://www.coe.int/t/dg4/linguistic/Ruschlikon1991_en.pdf> (30 de noviembre de 2015).

Consejo de Europa (1996a). *Modern Languages: Learning, Teaching, Assessment. A Common European Framework of Reference. Draft 2 a Framework Proposal.* CC-LANG (95)5 rev IV. Estrasburgo: Consejo de Europa.

Consejo de Europa (1996b). *Users' Guide for Examiners.* Estrasburgo: Consejo de Europa.

Consejo de Europa (1997), *European Language Portfolio: Proposals for Development.* CC-LANG (97)1. Estrasburgo: Consejo de Europa.

Consejo de Europa (2001). *Common European Framework for Languajes: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press. En: <http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf> (30 de noviembre de 2015).

Consejo de Europa (2002a), *Language Examining and Test Development.* Estrasburgo: ALTE - Consejo de Europa.

Consejo de Europa (2002b). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación.* Madrid: MECD y Anaya. En: <http://cvc.cervantes.es/obref/marco> (30 de noviembre de 2015).

Consejo de Europa (2005). *Reference Level Descriptions for National and Regional Languages (RLD), Guide for the Production of RLD (Versión 2)*, Estrasburgo: Language Policy División, DG IV. En: <https://www.coe.int/t/dg4/linguistic/ Source/DNR_Guide_EN.pdf > (30 de noviembre de 2015.

Consejo de Europa (2008). *Producciones orales que ilustran los 6 niveles del* Marco común europeo de referencia para las lenguas (DVD). Estrasburgo: Consejo de Europa. En: <http://www.ciep.fr/es/publicaciones-y-cd-roms-dedicados-a-evaluacion-y-a-certificacion/dvd-producciones-orales-ilustran-los-6-niveles-del-marco-comun-europeo-referencia-para-las-lenguas> (30 de noviembre de 2015).

Consejo de Europa (2009a). *Relating Language Examination to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A manual.* Estrasburgo: Consejo de Europa.

En: <http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf> (30 de noviembre de 2015).

Consejo de Europa (2009b). *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Estrasburgo: Consejo de Europa. En: <http://www.coe.int/t/dg4/linguistic/Source/ManualForeword ContSect A_2009_en.pdf> (30 de noviembre de 2015).

Consejo de Europa (2011). *Manual for Language Test Development and Examining for Use with the CEFR. Produced by ALTE on Behalf of the Language Policy Division., Council of Europe*. Estrasburgo: Consejo de Europa. En: <http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf> (30 de noviembre de 2015).

Consejo para la Cooperación Cultural (2002). *European Language Portfolio (ELP). Rules for the Accreditation of ELP Models (rev. 2)*. Estrasburgo: Consejo de Europa. En: <http://archivio.pubblica.istruzione.it/argomenti/ portfolio/allegati/regole_accredi_inglese.rtf > (30 de noviembre de 2015).

Consejo para la Cooperación Cultural (2004). *European Language Portfolio (ELP). Principles and Guidelines with Added Explanatory Notes (version 1.0)*. Estrasburgo: Consejo de Europa. En: <http://www.coe.int/t/dg4/ Linguistic/Source/Guidelines_EN.pdf> (30 de noviembre de 2015).

Cooper, William H. (1981). "Ubiquitous halo". *Psychological Bulletin,* 90, 218-244.

Corrigan, Michael (2007). *Seminar to Calibrate Examples of Spoken Performance. Università per Stranieri di Perugia, CVCL (Centro per la Valutazione e la Certificazione Linguistica) Perugia, 17–18 de diciembre de 2005. Report on the Analysis of Rating Data*. Estrasburgo: Consejo de Europa. En: <http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key_ reference/Report_Perugia05_EN.pdf> (30 de noviembre de 2015).

Coste, Daniel; Courtillon, Janine; Ferenczi, Victor, Martins-Baltar, Michel; Papo, Eliane y Rulet, Eddy (1976). *Un niveau-seuil*, París: Hatier.

Cronbach, Lee J. (1947). "Test reliability: its meaning and determination", *Psychometrika*, 12 (1), 1-16. En: < http://hbanaszak.mjr.uw.edu.pl/TempTxt/Amos%20FAQ's%20--%20TOC_files/Cronbach_1947_TestReliabilityItsMeaningandDetermination.pdf>.

Cronbach, Lee J. (1951). "Coefficient alpha and the internal satucture of tests", *Psychometrika*, 16 (3), 297-334). En: < http://kttm.hoasen.edu.vn/sites/default/files/2011/12/22/cronbach_1951_coefficient_alpha.pdf>.

Cronbach, Lee J. (1990). *Essentials of Psychological Testing* (5ª ed.). Nueva York: Harper & Row.

Cronbach, Lee J.; Gleser, Goldine C.; Nanda, Harinder y Rajaratnam, Nageswari (1972). *The Dependability of Behavioral Measurement: Theory of Generalizability for Scores and Profiles*, Nueva York, NY: Wiley.

Cronbach, Lee J.; Rajaratnam, Nageswari y Gleser, Goldine C. (1963). "Theory of Generalizability: A liberalization of reliability theory". *The British Journal of Statiscal Psychology*, 16 (2), 137-163.

Cursos Internacionales de la Universidad de Salamanca e Instituto Cervantes (2015). "Curso para correctores de pruebas de expresión e interacción escritas DELE". Curso presencial impartido en Salamanca, del 4 al 6 de febrero de 2015.

Cuxart, Anna; Martí, Manuel y Ferrer, Ferran (1987). "Algunos factores que inciden en el rendimiento y la evaluación en los alumnos de las pruebas de aptitud de acceso a la universidad". *Revista de Educación*, 314, 63-88. En: <http://www.mecd.gob.es/dctm/revista-de-educacion/articulosre314/re3140400462.pdf?documentId=0901e72b81272c3d> (30 de noviembre de 2015).

Ebel, Robert L. (1951). "Estimation of the reliability of ratings". *Psichometrika*, 16, 407-424.

Eckes, Thomas (2004). "Beurteilerübereinstimmung und Beurteilerstrenge: Eine Multi-facetten-Rasch-analyse von Leistungsbeurteilungen im «Test Deutch als Fremdsprache» (TestDaF) [Rater agreement and rater severity: A many-facet Rasch analysis of performance assessments in the «Test of German as a Foreign Language (TestDaF) »]". *Diagnostica*, 50, 65-77.

Eckes, Thomas (2005). "Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis". *Language Assessment Quaterly*, 2, 197-221.

Eckes, Thomas (2008). "Assuring the quality of TestDaF examinations: A psychometric modeling approach", en: Taylor, Lynda y Weir, Cyril J. (eds.), *Multilingualism and Assessment: Achieving Transparency, Assuring Quality, Sustaining Diversity – Proceedeings of the ALTE Berlin Conference May 2005*.

Cambridge: Cambridge University Press, 157-178

Eckes, Thomas (2009). "Many-facet Rasch measurement", en: Takala, Sauli (ed.), *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Section H). Estrasburgo: Consejo de Europa.

Eckes, Thomas (2010). "The TestDaF implementation of the SOPI: Design, analysis, and evaluation of a semi-direct speaking test", en: Araujo, Luisa (ed.), *Computer-based assessment (CBA) of Foreign Language Speaking Skills*. Luxemburgo: Oficina de Publicaciones de la Unión Europea, 63-83

Eckes, Thomas (2011). *Introduction to Many-facet Rasch Measurement. Analyzing and Evaluating Rater-Mediated Assessments*. Frankfurt: Peter Lang.

Ek, Jan Ate van (1975). *The Threshold Level in a European Unit/Credit System for Modern Lenguage Learning by Adults*. Estrasburgo: Consejo de Europa.

Ek, Jan Ate van (1986). *Objectives for Foreign Language Learning* (Vol I.). Estrasburgo: Consejo de Europa.

Ek, Jan Ate van y Alexander, L.G.; con la colaboración de Fitzpatrick, M.A. (1977/1980). *Waystage: an Intermediary Objective below Threshold Level in a European Unit/Credit System of Modern Language Learning by Adults*. Estrasburgo: Consejo de Europa, 1977 / *Waystage English*. Oxford: Pergamon Press (con el auspicio del Consejo de Europa), 1980.

EK, Jan Ate van y TRIM, John Leslie Melville (1991/1998). *Threshold Level 1990*. Cambridge: Cambridge University Press, 1991 / (edición revisada y corregida). Cambridge: Cambridge University Press, 1998.

Ek, Jan Ate van y Trim, John Leslie Melville (1991/1998). *Waystage 1990*, Cambridge: Cambridge University Press, 1991 / (edición revisada y corregida). Cambridge: Cambridge University Press, 1998. En: <https://www.coe.int/t/dg4/linguistic/Waystage_CUP.pdf> (30 de noviembre de 2015).

Ek, Jan Ate van y TRIM, John Leslie Melville (1996/2001), *Vantage Level*, Cambridge: Cambridge University Press, 1996 / (edición revisada y corregida). *Vantage*. Cambridge: Cambridge University Press, 2001. En: <https://www.coe.int/t/dg4/linguistic/Vantage_CUP.pdf> (30 de noviembre de 2015).

EK, Jan Ate van y Trim, John Leslie Melville (1997/2000), *Vantage Level*, Estrasburgo, Consejo de Europa, 1997 / Cambridge: Cambridge University Press, 2000).

Embretson, Susan E. y Hershberger, Scott L. (1999). *The New Rules of Measurement*. Mahwah, NJ: LEA.

Embretson, Susan E. y McCollam, Karen M. Schmidt (2000). "Psychometric approaches to understanding and measuring intelligence", en: Sternberg, Robert J. (ed.). *Handbook of Intelligence*. Cambridge: Cambridge University Press, 423-444 En: <https://smartech.gatech.edu/handle/1853/34546> (30 de noviembre de 2015).

Embretson, Susan E. y Reise, Steven Paul (2000) *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum Associates.

Engelhard, George (2002). "Monitoring raters in performance assessments", en: Tindal, Gerald y Haladyna, Thomas M. (eds.), *Large-scale Assessment Programs for all Students: Validity, Technical Adequacy, and Implementation.* Mahwah, NJ: Eribaum Associates, 261-287.

Engelhard, George (2008). "Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken". *Measurement: Interdisciplinary Research and Perspective,* 6, 155-189.

Engelhard, George (2013). *Invariant Measurement. Using Rasch Models in the Social, Behavioral, and Health Sciences*. Londres: Routledge.

Engelhard, George y Myford, Carol M. (2003). *Monitoring Faculty Consultant Performance in the Advanced Placement English Literature and Composition Program with a Many-Facet Rasch Model*, Nueva York, NY: College Entrance Examination Board. En: <http://onlinelibrary.wiley.com/store/10.1002/j.2333-8504.2003.tb01893.x/asset/ets201893.pdf?v=1&t=iekf18pq&s=7e003dd0a78b3fe38d30391e69c980be4009bb29> (30 de noviembre de 2015).

Farrokhi, Farahman y Esfandiari, Rajab (2011). "A Many-facet Rasch model to detect halo effect in three types of raters". *Theory and Practice in Language Studie,*. 1 (11), 1531-1540.

Fernández, Sonsoles (2003). *Propuesta curricular y Marco Común Europeo de Referencia*. Madrid: Consejería de Educación de Portugal/Edinumen.

Figueras, Neus (2004). "Estándares y calidad en la elaboración y administración de pruebas y exámenes. Criterios mínimos para el reconocimiento y la comparabilidad", en: *III Congreso de la Lengua Española (La certificación de la competencia lingüística en español como lengua extranjera. Hacia un enfoque hispánico del sistema)* celebrado en Rosario (Argentina) del 17 al 20 de noviembre de 2004. En: <http://www.congresosdelalengua.es/rosario/ponencias/ internacional/figueras_n.htm> (30 de noviembre de 2015).

Figueras, Neus (2008). "El MCER, más allá de la polémica". *MarcoELE*, 7, 26-35. En: <http://marcoele.com/descargas/evaluacion/03.figueras.pdf> (30 de noviembre de 2015).

Fischer, Gerhard H. (1973). "The linear logistic test model as an instrument in educational research". *Acta Psychologica,* 37 (6), 359-374. En: <http://www.researchgate.net/publication/247924556_The_linear_logistic_test_model_as_an_instrument_in_educational_research._Acta_Psychologica_37_359-374> (30 de noviembre de 2015).

Fischer, Gerhard H. (1995a). "Derivation of the Rasch model", en: Gerhard H. Fischer y Molenaar, Ivo W. (eds.), *Rasch Models: Foundations, Recent Developments, and Applications*. Nueva York: Springer, 15-38.

Fischer, Gerhard H. (1995b). "The linear logistic test model", en: Fischer, Gerhard H. y Molenaar, Ivo W. (eds.), *Rasch Models: Foundations, Recent Developments, and Applications*. Nueva York, NY: Springer, 131-155.

Fischer, Gerhard H. (2007). "Rasch models", en: Rao, Calyampudi Radhakrishna y Sinharay, Sandip (eds.), *Psychometrics (Handbook of Statistics 26*. Amsterdam: Elsevier Science B.V., 515-585.

Fisicaro, Sebastiano A. y Lance, Charles E. (1990). "Implications of three causal models for the measurement of halo error". *Applied Psychological Measurement,* 14, 419-429.

Fisicaro, Sebastiano A. y Vance, Robert J. (1994). "Comments on the Measurement of Halo". *Educational and Psychological Measurement*, 54 (2), 366-371.

Fitzpatrick, Robert y Morrison, Edward J. (1971). "Performance and product evaluation", en: Thorndike, Robert L. (ed.), *Educational Measurement* (2ª ed.). Washington, DC: American Council on Education, 237-270

Frederikse, John R., y Collins, Allan (1989). "A systems approach to educational testing". *Educational Researcher,* 18 (9), 27-32.

Fulcher, Glenn (1996), "Does thick description lead to smart tests? A data-based approach to rating scale construction". *Language Testing*, 13 (2), 208-238.

García, Álvaro (2002). "Bases comunes para una Europa plurilingüe: Marco común europeo de referencia para las lenguas", en: Instituto Cervantes, *Anuario del Instituto Cervantes 2002. El español en el mundo*, Barcelona: Plaza y Janés, Círculo de Lectores e Instituto Cervantes, 13-34.

Geranpayeh, Ardeshir (2001a). CB BULATS: "Examining the reliability of a computer based test using test- retest method". *Research Notes* 5, 14-16.

Gipps, Caroline V. (1994). *Beyond Testing*, Londres: Falmer Press.

Gleser, Goldine C., Cronbach, Lee J. y Rajaratnam, Nageswari (1965). "Generability of scores influenced by multiple sources of variance", *Psychometrika*, 30, 395-418.

Grossman, Michele y Wood, Wendy (1993). "Sex differences in emotional intensity: A social role explanation". *Journal of Personality and Social Psychology,* 65, 1010-1022. En: <http://dornsife.usc.edu/assets/sites/545/docs/Wendy_Wood_Research_Articles/Gender_Differences_in_Social_Behavior/Grossman.Wood.1993_Sex_differences_in_emotional_intensity.pdf> (30 de noviembre de 2015).

Guilford, Joy Paul (1954). *Psychometric Methods*. New York, NY: McGraw-Hill.

Gulliksen, Harold (1950). *Theory of mental tests*. Nueva York: Wiley

Gutiérrez, Salvador (2008). "Del arte gramatical a la competencia comunicativa". Discurso leído el día 24 de febrero de 2008 en su recepción pública como miembro de la Rean Academia Española. Madrid. (<http://www.rae.es/sites/default/files/Discurso_Ingreso_Salvador_Gutierrez.pdf> (30 de noviembre de 2015).

Haertel, Edward H. y Linn, Robert L. (1996). "Comparability", en: Gary W. Phillips (ed.), *Technical Issues in Large-scale Performance Assessment* (NCES 96-802). Washington, DC: National Center for Education Statistics, 59-78.

Hambleton, Ronald K.; Swaminathan, Hariharan y Rogers, H. Jane (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Harasym, Peter H.; Woloschuk, Wayne y Cunning, Leslie (2008). "Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs". *Advances in Health Sciences Education: Theory and Practice*, 13, 617-632.

Herzog, Martha (2011). *An Overview of the History of the ILR Language Proficiency Skill Level Descriptions and Scale*. En: <http://www.govtilr.org/Skills/IRL%20Scale%20History.htm> (30 de noviembre de 2015).

Hoyt, Cyril (1941). "Test reliability obtained by analysis of variance". *Psychometrika*, 6, 153-160.

Huhta, Ari; Luoma, Sari; Oscarson, Mats; Sajafaara, Kai; Takala, Sauli y Teasdale, Alex (2002). "DIALANG: A diagnostic language assessment system for adult learners", en: Alderson, J. Charles (ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies*. Estrasburgo: Consejo de Europa.

Huot, Brian (1990). "The literature of direct writing assessment: Major concerns and prevailing trends". *Review of Educational Research,* 60 (2), 237-263.

Hymes, D. (1972). "On communicative competence", en: Pride, John Bernard y Holmes, Janet (eds.), *Sociolinguistics*. Harmondsworth: Penguin, 269-293. Existe traducción al español del artículo con el título "Acerca de la competencia comunicativa", en: Llobera, Miquel (coord.) (1995), *La competencia comunicativa*. Madrid: Edelsa, 27-46.

Instituto Cervantes (1994). *Plan curricular del Instituto Cervantes*. Madrid: Instituto Cervantes.

Instituto Cervantes (2007). *Plan curricular del Instituto Cervantes. Niveles de referencia para el español.* Madrid: Instituto Cervantes - Biblioteca Nueva.

Instituto Cervantes (2012a). *Prueba de Expresión e interacción escritas* (tareas y hoja de respuestas). Viernes 25 de mayo de 2012. (Forma de examen no publicada).

Instituto Cervantes (2012b). *Prueba de Expresión e interacción escritas* (tareas y hoja de respuestas). Sábado 26 de mayo de 2012. (Forma de examen no publicada).

Instituto Cervantes (2013): *Guía del examen Diploma de Español Nivel A2*. En: <http://diplomas.cervantes.es/informacion/guias/materiales/a2/guia_examen_dele_a2.pdf> (30 de noviembre de 2015).

Instituto Cervantes (s.f. a). *Breve manual de uso de la aplicación HARES*. (Publicación de uso interno de marzo de 2015).

Instituto Cervantes (s.f. b). *Nivel A2. Explicación y ejemplo del examen*. En: <http://diplomas.cervantes.es/sites/default/files/modelo_examen_a2_1.pdf> (30 de noviembre de 2015).

Instituto Cervantes (s.f.). *Diccionario de términos claves de ELE*. Centro Virtual Cervantes. En: <http://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ ele/> (30 de noviembre de 2015).

Instrumento de Adhesión de España al Estatuto del Consejo de Europa (número 001), hecho en Londres el 5 de mayo de 1949 (1978). En: <

http://www.exteriores.gob.es/Portal/es/PoliticaExteriorCooperacion/ConsejoDeEuropa/Documents/Acta%20de%20Adhesi%C3%B3n%20de%20Espa%C3%B1a%20al%20Consejo%20de%20Europa.pdf> (30 de noviembre de 2015).

John, H. A. L. de Jong. (2001). "Procedures for relating test scores to Council of Europe framework". Ponencia presentada en la conferencia de ALTE celebrada en Barcelona en julio de 2001 titulada *European Language Testing in a Global Context*. Una selección de algunas de las ponencias presentadas puede consultarse en Milanovic, Michael y Weir, Cyril (eds.). (2004). *European Language Testing in a Global Context*. Cambridge: Cambridge University Press.

Johnson, Robert L.; Penny, James A. y Gordon, Belita (2009). *Assessing Performance: Designing, Scoring, and Validating Performance Tasks*. New York, NY: Guilford Press.

Jones, Neil (2002). "Relating the ALTE framework to the common european framework of reference", en: Alderson J. Charles (ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies*. Estrasburgo: Consejo de Europa.

Jones, Neil (2009). "A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard-setting", en: Figueras, Neus y Noijons, José (eds.). *Linking to the CEFR Levels: Research Perspectives*. Arnhem: Cito, Institute for Educational Measurement, Council of Europe y European Association for Language Testing and Assessment (EALTA), 35-43. En: <http://www.coe.int/t/dg4/linguistic/EALTA_Publicatie Colloquium2009.pdf> (30 de noviembre de 2015).

Jones, Neil (2014). *Multilingual Frameworks: the Construction and Use of Multilingual Proficiency Frameworks*. Cambridge: Cambridge University Press.

King, Larry M., Hunter, John E. y Schmidt, Frank L. (1980). "Halo in a multidimensional forced-choice performance evaluation scale". *Journal of Applied Psychology*, 65 (5), 507-516.

Kirk, Roger E. (1995). *Experimental Design: Procederes for the Behavioral Sciences* (3ª ed.). Pacific Grove, CA: Brooks/Cole.

Kirsch, Irwin S., y Mosenthal, Peter B. (1995), "Interpreting the IEA reading literacy scales", en Binkley, Marilyn; Rust, Keith y Winglee, Marianne (eds.). *Methodological issues in comparative educational studies: The case of the IEA reading literacy study*. Washington D.C.: Department of Education. National Center for Education Statistics, 135-192.

Kondo-Brown, Kimi (2002). "An analysis of rater bias with FACETS in measuring Japanese L2 writing performance". *Language Testing*, 19, 1-29.

Kubinger, Klaus D. (2009). "Applications of the linear logistic test model in psychometric research". *Educational and Psychological Measurement,* 69, 232-244.

Landy, Frank J. y Farr, James L. (1980). "Performance rating". *Psychological Bulletin*, 87 (1), 72-107.

Landy, Frank J. y Farr, James L. (1983). *The Measurement of Work Performance: Methods, theory, and applications*. San Diego, CA: Academic Press.

Lane, Suzanne y Stone, Clement A. (2006). "Performance assessment", en: Brennan, Robert L. (ed.). *Educational Measurement*, (4ª edición). Westport, CT: American Council on Education and Praeger, 387–431.

Likert, Rensis (1932). *A Technique for the Measurement of Attitudes*. Nueva York, NY: Archives of Psychology.

Linacre, John Michael (1989). *Many-facet Rasch Measurement*. Chicago: MESA Press.

Linacre, John Michael (2000). "Comparing «Partial Credit Models» (PCM) and «Rating Scale Models» (RSM)". *Rasch Measurement Transactions*, 14 (3), 768.

Linacre, John Michael (2002a). "Optimizing rating scale category effectiveness". *Journal of Applied Measurement*, 3, 85-106.

Linacre, John Michael (2002b). "What do infit and ourfit, mean-square and standardized mean?". *Rasch Measurement Transactions*, 16 (2), 878.

Linacre, John Michael (2003). "Size vs. significance: Infit and ourfit mean-square and standardized chi-square fit statistic". *Rasch Measurement Transactions*, 17 (1), 918.

Linacre, John Michael (2004). "Estimation methods for Rasch measures", en: Smith, Everett V. y Smith, Richard M. (eds.). *Introduction to Rasch Measurement*, Maple Grove, MN: JAM Press, 25-47.

Linacre, John Michael (2005). "Standar errors: Means, measures, origins and anchor values". *Rasch Measurement Transactions*, 19 (3), 1030.

Linacre, John Michael (2006). "Demarcating category intervals". *Rasch Measurement Transactions*, 19 (4), 1041-1043.

Linacre, John Michael (2010a). *A user's guide to FACETS: Rasch-model Computer*

Programs. Chicago: Winsteps.com.

Linacre, John Michael (2010b). "Transitional categories and usefully disordered thresholds". *Online Educational Research Journal,* 1 (3), 1-10. En: <http://www.oerj.org/View?action=viewPDF&paper=2> (30 de noviembre de 2015).

Linacre, John Michael (2012). *Facets Computer Program for Many-facet Rasch Measurement, versión 3.70.0*. Beaverton, OR: Winsteps.com.

Linacre, John Michael y Wright, Benjamín D. (2002). "Construction of measures from Many-facet data". *Journal of Applied Measurement*, 3 (4), 484-509.

Lindquist, Everet Franklin (1953). *Design and Analysis of Experiments in Education and Psychology*. Boston, MA: Houghton Mifflin.

Linn, Robert L.; Baker, Eva L. y Dunbar, Stephen B. (1991). "Complex performance assessment: Expectations and validation criteria". *Educational Researcher,* 20 (8), 15-21.

Little, David, Goullier, Francis y Hughes, Gareth. (2011). *The European Language Portfolio: the Story so Far (1991-2011)*. Estrasburgo: Consejo de Europa.

Llobera, Miquel (coord.) (1995). *Competencia comunicativa: documentos básicos en la enseñanza de lenguas extranjeras*. Madrid: Edelsa.

Llorián, Susana (2007). *Entender y utilizar el* Marco común europeo de referencia *desde el punto de vista del profesor de lenguas*. Madrid: Santillana-Universidad de Salamanca.

Lunz, Mary E., Wright, Benjamin D., and Linacre, John Michael (1990) "Measuring the impact of judge severity on examination scores". *Applied Measurement in Education*, 3 (4), 331-345.

Martínez, Mª Rosario (1995). *Psicometría: teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Martínez, Mª Rosario (2010): "La evaluación del desempeño", *Papeles del Psicólogo*, 31 (1), 85-96.

Mártínez, Mª Rosario; Hernández, Mª José y Hernández, Mª Victoria (2006). *Psicometría*. Madrid: Alianza Editorial.

Martyniuk, Waldemar y Noijons, José (2007). "Executive summary of results of a survey on The use of the CEFR at national level in the Council of

Europe Member States". En: Goullier, Francis (relator), *The* Common European Framework of Reference for Languages *(CEFR) and the development of language policies: challenges and responsibilities*. Estrasburgo: Consejo de Europa.

Marzano, Robert J., Pickering, Debra, y McTighe, Jay (1993). *Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model.* Alexandria, VA: Association form Supervision and Curriculum Development.

Masters, Geoffery N. (1982). "A Rasch model for partial credit scoring". *Psychometrika*, 47, 149-174.

Masters, Geoffery N. y Wright, Benjamin D. (1997). "The partial credit model". En: Linden, Wim J. van der y Hambleton, Ronald K. (eds.), *Handbook of Modern Item Response Theory*. Nueva York, NY: Springer-Verlag, 111-121.

McManus, Ian Christopher, Thompson, Margaret y Mollon, Jennifer (2006). "Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP (UK) clinical examination (PACES) using multi-facet Rasch modeling". *BMC Medical Education*, 6, 1272-1294. En: <http://link.springer.com/article/10.1186%2F1472-6920-6-42#page-2> (30 de noviembre de 2015).

McNamara, Timothy Francis (1996). *Measuring Second Language Performance.* London: Longman.

Messick, Samuel J. (1989). "Validity", en: Linn, Robert L. (ed.). *Educational measurement* (3ª ed.). New York, NY: American Council on Education and Mcmillan, 13-104.

Messick, Samuel J. (1994). "The interplay of evidence and consequences in the validation of performance assessments". *Educational Researcher,* 23 (2), 13-23.

Milanovic, Michael, y Saville, Nick (eds.) (1996), *Performance testing, cognition and assessment*, Cambridge: University of Cambridge Local Examinations Syndicate.

Miller, M. David y Crocker, Linda (1990). Validation methods for direct writing assessment. *Applied Measurement in Education,* 3 (3), 285-296.

Ministerio de Educación, Cultura y Deporte (España) (2003). *El portfolio europeo de las lenguas: guía para profesores y formadores de profesores*. Estrasburgo: Consejo de Europa. (Original en inglés en: Little, David y Perclová, Radka (2001). *The European Language Portfolio: guide for teachers and teacher*

trainers. Estrasburgo: Consejo de Eurpoa. En: <http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Publications/ELPguide_teacherstrainers_EN.pdf> (30 de noviembre de 2015).

Mullis, Ina V. S. (1980). *Using the primary trait-system for evaluating writing*. Denver, CO: National Assessment of Education Progress, Education Commission of the States.

Mullis, Ina V. S. (1984). "Scoring direct writing assessments: What are the alternatives?". *Educational Measurement: Issues and Practice,* 3 (1), 16-18.

Muñiz, José (2000). *Teoría Clásica de los Tests*. Madrid: Pirámide.

Muñiz, José (2010). "Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems". *Papeles del Psicólogo*, 31(1), 57-66

Muñiz, José y Hambleton, Ronald K. (1992). "Medio siglo de teoría de respuesta al ítem". *Anuario de Psicología,* 52*,* 41-66.

Myford, Carol M. y Wolfe, Edward W. (2004a). "Detecting and measuring rater effects using many-facet Rasch measurement: Part I", en: Smith, Everett V. y Smith, Richard M. (eds.). *Introduction to Rasch measurement*. Maple Grove, MI: JAM Press, 460–517.

Myford, Carol M. y Wolfe, Edward W. (2004b). "Detecting and measuring rater effects using many-facet Rasch measurement: Part II", en: Smith, Everett V. y Smith, Richard M. (eds.). *Introduction to Rasch measurement*. Maple Grove, MI: JAM Press, 518-574.

Myford, Carol M. y Wolfe, Edward W. (2009). "Monitoring rater performance over time: A framework for detecting differencital accuracy and differential Scale category use". *Journal of Educational Measurement,* 46, 371–389.

Myford, Carol M., Marr, Diana B., y Linacre, John Michael (1996). *Reader Calibration and its Potential Role in Equating for the Test of Written English* (TOEFL Research Report 52). Princeton, NJ: Educational Testing Service. En: <https://www.ets.org/Media/Research/pdf/RR-95-40.pdf> (30 de noviembre de 2015).

Navarro, Pedro y Navarro, Román M. (2008). "Cómo aplicar los Niveles de Referencia a la elaboración de materiales didácticos: estudio sobre Pasaporte A1". *MarcoELE*, 6, 1-28.

Navas, María José (1994). "Teoría Clásica de los Tests versus Teoría de Respuesta al Item". *Psicológica* 15, 175-208. En:

<http://www.uned.es/490015/CV/TCTTRI94.pdf> (30 de noviembre de 2015).

North, Brian *(*1997). "The development of a common framework scale of descriptors of language proficiency based on a theory of measurement". Ponencia presentada en *Language Testing Research Colloquium* (LTRC, en las siglas en inglés) en 1996 en Tampere (Finlandia), en: Huhta, Ari; Kohonen, Viljo; Kurki-Suonio, Liisa y Luoma, Sari. *Current Developments and Alternatives in Language Assessment*. Jyvskyl: University of Jyvskyl, 423-449.

North, Brian (2008). "The CEFR levels and descriptive scales", en: Taylor, Linda y Weir Cyril J. (eds.), *Multilingualism and Assessment: Achieving Transparency, Assuring Quality, Sustaining Diversity—Proceedings of the ALTE Berlin Conference May 2005*. Cambridge: Cambridge University Press, 21-66

North, Brian y Schneider, Günther (1998). "Scaling descriptors for language proficiency scales". *Language Testing*, 15, 217-262.

North, Brian (1996/2000). *The development of a common framework scale of descriptors of language proficiency based on a theory of measurement*. Tesis doctoral. Thames Valley University, 1996 / Nueva York, NY: Peter Lang, 2000.

North, Brian y Jones, Neil (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling.* Estrasburgo: Consejo de Europa. En: <http://www.coe.int/t/dg4/linguistic/Publications_EN.asp> (30 de noviembre de 2015).

OCDE (2006). *PISA 2003. Manual de análisis de datos. Usuarios de SPSS®*. En: <http://www.mecd.gob.es/dctm/ievaluacion/internacional/pisamanualdatos.pdf?documentId=0901e72b80110555>. (Original en inglés en: OECD (2005). *PISA 2003. Data Analysis Manual: SPSS® Users*. En: <http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/35004299.pdf>).

Ostini, Remo, y Nering, Michael L. (2005). *Polytomous item responser theory*. Thousand Oaks, CA: Sage.

Park, Taejoon (2004). "An investigation of an ESL placement test of writing using Many-Facet Rasch measurement". *Papers in TESOL y Applied Linguistics*, 4, 1-21.

Parker, William R. (1957). *The National Interest and Foreingn Languages*. Washington D. C.: United States National Committee for UNESCO, Department of State.

Paz, María (1988). "La enseñanza del español en los Estados Unidos". *Cable*, 2, 26-34.

Pollitt, Alastair, y Murray, Neil L. (1996). "What raters really pay attention to", en: Milanovic, Michael, y Saville, Nick (eds.) (1996), *Performance testing, cognition and assessment*, Cambridge: University of Cambridge Local Examinations Syndicate, 74-91.

Popham, W. James (1990). *Problemas y técnicas de la evaluación educativa*. Madrid: Anaya.

Pride, J.B. and Holmes, J. (eds) (1972) *Sociolinguistics*. Harmondsworth: Penguin.

Prieto, Gerardo (2011). "Evaluación de la ejecución mediante el modelo Many-Facet Rasch Measurement". *Psicothema*, 23: 2, 233-238.

Prieto, Gerardo (2014). *Breve descripción de la metodología psicométrica empleada en el análisis de los exámenes del DELE*. Instituto Cervantes – Universidad de Salamanca. (Informe interno de marzo de 2014).

Prieto, Gerardo (2015a). *Análisis de la severidad de los calificadores de los exámenes de expresión escrita del DELE realizados en 2014*. Instituto Cervantes – Universidad de Salamanca. (Informe interno de marzo de 2015).

Prieto, Gerardo (2015b). "Análisis de un test de desempeño en expression escrita mediante el modelo de MFRM". *Actualidades en Psicología,* 29 (119), 1-17.

Prieto, Gerardo e Dias, Angela (2003) "Uso del modelo de Rasch para poner en la misma escala las puntuaciones de distintos tests"**.** *Actualidades en Psicología,* 19, 106, 5-23. En: < http://www.revistas.ucr.ac.cr/index.php/actualidades/article/view/43/34> (30 de noviembre de 2015).

Prieto, Gerardo y Delgado, Ana R. (2003). "Análisis de un test mediante el modelo de Rasch", *Phicothema*, 15, 1, 94-100.

Prieto, Gerardo y Delgado, Ana R. (2010). "Fiabilidad y validez", *Papeles del Psicólogo*, 31, 1, 67-74.

Prieto, Gerardo y Nieto, Eloísa (2014). "Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement". *Psicológica,*

35, 363-375.

Puig, Fuensanta. (2008). "El proyecto DIALANG", en: Pastor Cesteros, Susana y Roca Marín, Santiago (eds.). *La evaluación en el aprendizaje y la enseñanza del español como lengua extranjera (LE) y segunda lengua (L2)*, Alicante: Universidad de Alicante, 76-79.

Raban, Sandra (2008). *Examining the World. A History of the University of Cambridge*. Cambridge: Cambridge University Press.

Rasch, Georg (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhague: Danish Institute for Educational Research, 1960 / Chicago, IL: University of Chicago Press, 1980.

Real Academia Española (2014). *Diccionario de la lengua española (23.ª ed.)*. Madrid: Espasa Libros. S. L. U.

Robbins, Stephen P. (1989). *Organizational behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Rodríguez, José Luis y Tejedor, Francisco Javier (1996). *Evaluación educativa: 1. Evaluación de los aprendizajes de los alumnos*. Salamanca: Ediciones Universidad de Salamanca.

Rost, Jürgen (1990). "Rasch models in latent classes: An Swender of two approaches to item analysis". *Applied Psychological Measurement*, 14 (3), 271-282.

Rost, Jürgen (2001). "The growing family of Rasch models", en: Boomsma, Anne; Duijn, Marijtje A. J. van y Snijders, Tom A. B. (eds.), *Essays on item response theory*. Nueva York, NY: Springer, 25-42.

Rost, Jürgen (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2ª ed.) Berna: Huber.

Saal, Frank E.; Downey, Ronald G., y Lahey, Mary A. (1980). "Rating the ratings: Assessing the psychometric quality of rating data". *Psychological Bulletin*, 88 (2), 413-428.

Scarino, Angela (1996). "Issues in planning, describing and monitoring long-term progress in language learning", en Scarino, Angela (Ed.), *Equity in Languages other than English*. Perth: Australian Federation of Modern Language Teachers Associations, 68–75.

Scarino, Angela (1997). "Analysing the language of frameworks of outcomes for foreign language learning". In Voss, Peter (ed.*) Joining Voices. Conference Proceedings of the AFMLTA Eleventh National Languages Conference.*

Hobart: AFMLTA.

Schneider, Günther y Lenz, Peter (2001). *European Language Portfolio: Guide for developers*. Estrasburgo: Consejo de Europa. En: <http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Publications/ Developers_guide_EN.pdf > (30 de noviembre de 2015).

Schneider, Günther y North, Brian (1999). "In anderen Sprachen kann ich... Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung des fremdsprachlichen Kommunicationmsfähigkeit", informe del proyecto, Programa Nacional de Investigación 33. Berna: Consejo Nacional Suizo de Investigación Científica.

Schneider, Günther y North, Brian (2000). *Fremdsprachen können -was heisst das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit*. Coira/Zurich: Verlag Rüegger AG.

Schriesheim, Chester A., Kinicki, Angelo J., y Schriesheim, Janet F. (1979). "The effect of leniency on leader behavior descriptions". *Organizational Behavior and Human Performance,* 23, 1–29.

Shavelson, Richard J. y Webb, Noreen M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Slagter, Peter Jan (1979). *Un Nivel Umbral*. Estrasburgo: Consejo de Europa, En: <http://88.2.223.134/linguaglobe/index_files/Student_file/ Documentos%20Oficiales/nivel_umbral.pdf> (30 de noviembre de 2015).

Slagter, Peter Jan (1990). *Threshold 1990*. Estrasburgo: Publicaciones del Consejo de Europa. En: <http://88.2.223.134/linguaglobe/index_files/ Student_file/Documentos%20Oficiales/nivel_umbral.pdf> (30 de noviembre de 2015).

Smith, Richard M., Schumaker, Randall. E. y Bush, M. Joan (1998). "Using item mean squares to evaluate fit to the Rasch model". *Journal of Outcome Measurement,* 2 (1), 66-78.

Spearman, Charles (1904). "The proof and measurement of association between two things". *American Journal of Psychology*, 15 (1) , 72-101. En: <http://www.jstor.org/stable/1412159?seq=1#page_scan_tab_contents>.

Spearman, Charles (1907). "Demonstration of formulae for true measurement of correlation". *American Journal of Psychology*, 18 (2), 161-169. En: < https://archive.org/stream/jstor-1412408/1412408_djvu.txt>.

Spearman, Charles (1910). "Correlations calculated from faulty data. *British Journal of Psychology*", 3 (3), 271-295.

Spearman, Charles (1913). "Correlations of sums and differences. *British Journal of Psychology*", 5 (4), 417-426.

Stanley, Julian C. (1971). "Reliability". En: Thorndike, Robert L. (ed.). *Educational Measurement* (2ª ed.). Washington, DC: American Council on Education.

Stockford, Lee, y Bissell, H. W. (1949). "Factors involved in establishing a merit-rating scale". *Personnel*, 26, 94-118.

Stone, Gregory Ethan (2006). "Whose criterion standard is it anyway?". *Journal of Applied Measurement,* 7 (2), 160-169.

Stray, Christopher (2001). "The Shift from Oral to Written Examination: Cambridge and Oxford 1700–1900". *Assessment in Education: Principles, Policy and Practice*, 8 (1), 33-50.

Sudweeks, Richard R.; Reeve, Suzanne y Bradshaw, William S. (2005). "A comparasion of generalizability theory and many-facet Rasch measurement in an analysis of college spphomore writing". *Assessing Writing*, 9, 239-261.

Swender, Elvira. (2001). "La enseñanza de ELE en los EE.UU.: cantidad y calidad", *II Congreso de Valladolid: paneles y ponencias. El activo del español: La industria del español como lengua extranjera.* Valladolid: España. En: <http://congresosdelalengua.es/valladolid/ponencias/activo_del_espanol/1_la_industria_del_espanol/swender_e.htm> (30 de noviembre de 2015).

Tesio, Luigi; Simone, Anna; Grzeda Marius T.; Ponzio, Michela; Dati, Gabriele, Zaratin, Paola, Perucca, Laura y Battaglia, Mario A. (2015). "Funding Medical Research Projects: Taking into Account Referees' Severity and Consistency through Many-Faceted Rasch Modeling of Projects' Scores". *Journal of Applied Measurement*, 16 2), 129-152.

Thorndike, Edward L. (1920). "A constant error in psychological ratings". *Journal of Applied Psychology* 4, 25-29. En: <http://www.romolocapuano.com/wp-content/uploads/2013/05/Thorndike.pdf> (30 de noviembre de 2015).

Thorndike, Robert L. (1951). "Reliability", en: Lindquist, Everett Franklin (ed.), *Educational Measurement*, Washington, DC: American Council on Education.

Thorndike, Robert L. y Hagen, Elizabeth P. (1977). "Measurement and evaluation in psychology and education", Nueva York, NY: John Wiley and Sons.

Trim, John Leslie Melville (1980). *Developing a Unit/Credit scheme of adult language learning*. Oxford: Pergamon.

Trim, John Leslie Melville (1981). Project 4 - *Modern Languages Programme 1971-81: Report Presented by CDCC Project Group 4 with a Resume*. Estrasburgo: Consejo de Europa.

Trim, John Leslie Melville (1997). *Language Learning for European Citizenship. Final Report of the Project Group Activities (1989-1996*), Estrasburgo: Consejo de Europa.

Trim, John Leslie Melville (ed.) (1978/1979): *Some Possible Lines of Development of an Overall Structure for a European Unit / Credit Scheme for Foreign Language Learning by Adults*. Estrasburgo: Consejo de Europa, 1978 / *Des voies possibles pour l'élaboration d'une structure générale d'un système européen d'unités capitalisables pour l'apprentissage des langues vivantes des adultes*). Estrasburgo: Consejo de Europa, 1979.

TRIM, John Leslie Melville (ed.), Bailly, Sophie; Devitt, Sean; Gremmo, Marie-José; Heyworth, Frank; Hopkins, Andy; Jones, Barry; Makosch, Mike; Riley, Philip y Stoks, Gé (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment. A Guide for Users*. Estrasburgo: Consejo de Europa.

Trim, John Leslie Melville (s.f.). Modern *Languages in the Council of Europe 1954-1997. International co-operation in support of lifelong language learning for effective communication, mutual cultural enrichment and democratic citizenship in Europe.* <http://www.coe.int/t/dg4/linguistic/Publications_en.asp> (30 de noviembre de 2015).

Trim, John Leslie Melville; Richterich, R.; Ek, Jan Ate van y Wilkins, David A. (1980). *Systems development in adult language learning,* Oxford: Pergamon.

Tyndall, Belle y Kenyon, Dorry Mann (1996). "Validation of a new holistic rating scale using Rasch multi-faceted analysis", en: Cumming, Alister H. y Berwick, Richard (eds.), *Validation in language testing*. Clevedon: Multilingual Matters, 39-57.

Upshur, John A. y Turner, Carolyn E. (1995). "Constructing rating scales for

second language tests", en: *English Language Tea- ching Journal*, 49 (1), 3-12.

Van der Linden, Wim J., y Hambleton, Ronald K. (1997). *Handbook of modern ítem response theory*. Nueva York, NY: Springer.

Verhelst, Norman D. y Glas, Cees A. W. (1995). "The generalized one parameter model: OPLM", en: Fischer, Gerhard H. y Molenaar, Ivo W. (eds.), *Rasch Models: Foundations, Recent Developments, and Applications*. Nueva York, NY: Springer-Verlag, 215-238.

Verhelst, Norman D.; Glas, Cees A. W. y Verstralen, Huub H. F. M. (1995). *OPLM: One Parameter Logistic Model. Computer program and manual*. Arnhem: CITO.

Watts, Frances y García, Amparo (eds.) (2006). *La evaluación compartida: investigación multidisciplinar*. Valencia: Editorial de la UPV. En: <http://www.upv.es/gie/LinkedDocuments/descargar%20libro.pdf> (30 de noviembre de 2015).

Weigle, Sara Cushing (2002). *Assessing writing*. Cambridge: Cambridge University Pres.

Welch, Catherine (2006). "Item and prompt development in performance testing", en Downing, Steven y Haladyna, Thomas M. (eds.), *Handbook of test development*. Mah- wah, NJ: Erlbaum, 303–327.

Wells, Frederic Lyman (1907). "A statistical study of literary merit with remarks on some new phases of the method". *Archives of Psychology* 7, 5-30.

Wikins, David A. (1978). "Proposal for Levels Definition", en: Trim, John Leslie Melville (ed.). *Some Possible Lines of Development of an Overall Structure for a European Unit / Credit Scheme for Foreign Language Learning by Adults*. Estrasburgo: Consejo de Europa.

Wolfe, Edward W. (2004). "Identifying rater effects using latent trait models". *Psychology Science*, 46, 35-51.

Wolfe, Edward W. (2009). "Item and rater analysis of constructed response ítems via the multi-faceted Rasch model". *Journal of Applied Measurement*, 10, 335-447.

Wolfe, Edward W. y Dobria, Lidia (2008). "Applications of the multifaceted Rasch model", en: Osborne, Jason W. (ed.), *Best practices in quantitative methods*. Los Angeles, CA: Sage, 71-85.

Wright, B. D. y Linacre, John Michael (1994). "Reasonable mean-square fit values." *Rasch Measurement Transactions*, 8, 370.

Wright, Benjamín D. y Mok, M. M. C. (2004). "An overview of the family of Rasch measurement models", en: Smith, Everett V. y Smith, Richard M (eds.), *Introduction to Rasch measurement*, 1-24. Maple Grove, MN: JAM Press.

Wright, Benjamín D. y Stone, Mark H. (1979). *Best test design. Rasch measurement*. Chicago, IL: MESA Press.

Wright, Benjamín D., y Masters, Geoffery N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Yorozuya, Ryuichi y Oller, John W., Jr. (1980). "Oral proficiency scales: Construct validity and the halo effect". *Language Learning*, 30 (1), 135-153.