

Anejo A: Desarrollo de los descriptores de dominio de la lengua

Este anejo analiza los aspectos técnicos de la descripción de niveles lingüísticos alcanzados; en él se estudian los criterios para la formulación de descriptores para enumerar después las metodologías para el desarrollo de escalas y ofrecer una bibliografía comentada.

La formulación de descriptores

La experiencia de elaborar escalas de medición en la evaluación de lenguas, la teoría del escalonamiento en el campo más amplio de la psicología aplicada y las preferencias de los profesores cuando participan en procesos de consulta (por ejemplo: los esquemas de gradación de objetivos del Reino Unido y del proyecto suizo) sugieren el siguiente conjunto de orientaciones para el desarrollo de descriptores:

- **Formulación positiva:** Una característica habitual de las escalas de dominio de la lengua centradas en el examinador y de las escalas de valoración de exámenes para la formulación de enunciados en los niveles inferiores, es que están redactadas de forma negativa. Resulta más difícil formular el dominio de la lengua en niveles inferiores en función de lo que el alumno sabe hacer que en función de lo que no sabe hacer. Pero si los niveles de dominio de la lengua tienen que servir de objetivos, más que de simples instrumentos para seleccionar a los examinandos, entonces se requiere una formulación positiva. Hay ocasiones, sin embargo, en que es posible formular el mismo elemento de forma positiva o negativa, por ejemplo, con relación al alcance de la lengua (véase el cuadro A1).

Una complicación añadida al evitar la formulación negativa es que existen algunas características del dominio lingüístico comunicativo que no son acumulativas; cuantas menos haya, por tanto, mejor. El ejemplo más evidente es lo que a veces se denomina independencia, esto es, el grado en que el alumno depende de a) el reajuste del discurso por parte del interlocutor, b) la posibilidad de pedir aclaración y c) la posibilidad de conseguir ayuda a la hora de formular lo que quiere decir. A menudo estos puntos se pueden tratar en condiciones añadidas a los descriptores especificados positivamente; por ejemplo:

Comprende generalmente el habla clara y normalizada sobre asuntos cotidianos que va dirigida a él, siempre que pueda pedir que le repitan o le vuelvan a formular de vez en cuando lo dicho.

Comprende lo que se le dice de forma clara, lenta y directa en conversaciones sencillas y cotidianas; puede llegar a comprender si el interlocutor se toma la molestia.

o bien:

Interactúa con razonable facilidad en situaciones estructuradas y conversaciones breves siempre que el interlocutor ayude, si es necesario.

Cuadro A1. Evaluación: criterios positivos y negativos

Positivo	Negativo
<p>Tiene un repertorio básico de elementos lingüísticos y de estrategias que le permiten abordar situaciones habituales predecibles. (Nivel 3 de <i>Eurocentres</i>: certificado)</p> <p>Tiene un repertorio lingüístico y de estrategias suficiente para enfrentar la mayoría de las necesidades habituales, pero generalmente requiere la búsqueda de palabras y aceptar la limitación del mensaje. (Nivel 3 de <i>Eurocentres</i>: cuadro del examinador).</p>	<p>Tiene un escaso repertorio lingüístico, que exige paráfrasis y búsqueda de palabras constantemente. (Nivel 3 de ESU).</p> <p>El dominio limitado de la lengua provoca frecuentes interrupciones y malentendidos cuando se enfrenta a situaciones poco habituales. (Nivel finlandés 2)</p> <p>La comunicación se interrumpe porque las restricciones lingüísticas interfieren en el mensaje. (Nivel 3 de ESU).</p>
<p>Vocabulario restringido a ciertas áreas, tales como objetos y lugares básicos y los parentescos más habituales. (Principiante ACTFL).</p>	<p>Tiene tan sólo un vocabulario limitado (Nivel holandés 1).</p> <p>La cantidad tan limitada de palabras y expresiones obstaculiza la comunicación de pensamientos e ideas. (Gotemburgo U).</p>
<p>Expresa y reconoce un conjunto de palabras y expresiones breves aprendidas de memoria (Trim 1978 Nivel 1).</p>	<p>Sólo es capaz de expresar listas y enumeraciones de enunciados convencionales (Principiante ACTFL).</p>
<p>Produce expresiones breves y habituales para satisfacer necesidades básicas concretas (en el área de saludos, información, etc.) (Elviri; Milán Nivel 1, 1986).</p>	<p>Tiene tan sólo un repertorio lingüístico muy básico; apenas muestra un dominio funcional de la lengua. (Nivel 1 de ESU).</p>

– **Precisión:** Los descriptores deben describir tareas concretas y, en su caso, grados concretos de destreza a la hora de realizar tareas. Subyacen, por tanto, dos ideas. En primer lugar, el escritor tiene que evitar vaguedades del tipo «utiliza una serie de estrategias adecuadas». ¿Qué se entiende por **estrategias**? ¿Adecuado a qué? ¿Cómo hay que interpretar *serie*? El problema de los descriptores imprecisos es que se leen muy bien, pero la aparente facilidad para aceptarlos puede ocultar el hecho de que cada cual lo interpreta de forma distinta. En segundo lugar, desde los años cuarenta se ha establecido el principio de que las distinciones entre las fases de una escala no deberían depender de la sustitución de un cuantificador como «algún» o «unos pocos» por «muchos» o «la mayoría», o de «bastante amplio» por «muy amplio», o «moderado» por «bueno» en el nivel inmediatamente superior, sino que tales distinciones deberían ser reales, no procesadas por palabras, y esto puede indicar que hay «lagunas» donde no se pueden realizar distinciones significativas ni concretas.

– **Claridad:** Los descriptores tienen que ser transparentes y no estar dominados por la jerga. Aparte de constituir una barrera para la comprensión, a veces ocurre que cuando a un descriptor aparentemente magnífico se le despoja de la jerga, resulta que decía muy poco. En

segundo lugar, los descriptores deben estar escritos con una sintaxis sencilla y tener una estructura explícita y lógica.

- **Brevedad:** Una de las escuelas de pensamiento se asocia con las escalas holísticas, sobre todo con las utilizadas en América y en Australia, e intenta elaborar un párrafo extenso que abarque de forma integradora las que se consideran características principales. Dichas escalas consiguen precisión mediante una lista integradora que -se pretende- transmita una imagen detallada de lo que los examinadores pueden reconocer como alumno típico de un nivel concreto, y por ello son fuentes muy valiosas de descripción. Dicho enfoque, sin embargo, tiene dos inconvenientes. En primer lugar, ningún individuo es realmente característico; las características detalladas coexisten de maneras distintas. En segundo lugar, no se puede hacer referencia de forma realista a un descriptor que abarque más de una oración compuesta con dos partes durante el proceso de evaluación; los profesores constantemente parecen preferir descriptores cortos, y en el proyecto de elaboración de los descriptores ilustrativos, solían rechazar o dividir los descriptores de más de veinticinco palabras (aproximadamente, dos líneas en letra de cuerpo normal).
- **Independencia:** Los descriptores breves tienen otras dos ventajas. En primer lugar, es más probable que describan un comportamiento del que se pueda decir: «Sí, esta persona sabe hacer esto». Por consiguiente, los descriptores breves y concretos se pueden utilizar como especificaciones independientes de criterios en listas de control y en cuestionarios para la evaluación continua que lleva a cabo el profesor y, en su caso, para la autoevaluación. Este tipo de integridad independiente es una señal de que el descriptor podría servir como objetivo, pues su sentido no sólo se deriva de la formulación de otros descriptores de la escala. Esto brinda una serie de oportunidades para su utilización en distintas formas de evaluación (véase el capítulo 9).

Los usuarios del Marco de referencia pueden tener presente y, en su caso, determinar:

- *Qué criterios de los presentados son más adecuados, y qué otros criterios se utilizan de forma explícita o implícita en su ámbito.*
- *Hasta qué punto es deseable y viable que las formaciones de su sistema cumplan criterios tales como los presentados aquí.*

Las metodologías del desarrollo de escalas

La existencia de una serie de niveles presupone que determinados elementos se pueden colocar en un nivel y no en otro y que las descripciones de un grado concreto de destreza pertenecen a un nivel y no a otro. Esto supone una forma de escalonamiento aplicada consistentemente. Hay varias formas posibles de asignar descripciones de dominio de la lengua a distintos niveles; los métodos actuales se pueden clasificar en tres grupos: métodos intuitivos, cualitativos y cuantitativos. La mayoría de las escalas de dominio de la lengua existentes, así como otros conjuntos de niveles, se han desarrollado mediante uno de los tres métodos intuitivos del primer grupo y los mejores métodos combinan los tres enfoques en un proceso complementario y acumulativo. Los métodos cualitativos requieren la preparación y la selección

intuitivas de material y la interpretación intuitiva de los resultados, mientras que los métodos cuantitativos deberían cuantificar cualitativamente el material comprobado previamente, y requieren una interpretación intuitiva de los resultados. Por tanto, al desarrollar los niveles comunes de referencia se utilizó una combinación de enfoques intuitivos, cualitativos y cuantitativos.

Si se utilizan los métodos cualitativos y cuantitativos, entonces hay dos puntos de partida posibles: descriptores o muestras de actuación.

Partir de los descriptores: Un punto de partida consiste en considerar lo que se desea describir, y después escribir, reunir o corregir borradores de descriptores de las categorías concretas como información de entrada de la fase cualitativa. Los métodos 4 y 9 -el primero y el último del grupo de los cualitativos que se muestran más adelante- son ejemplos de este enfoque, que es 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. La ventaja de partir de categorías y descriptores es que se puede definir una base teórica equilibrada.

Partir de muestras de actuación: La alternativa, que sólo puede utilizarse para desarrollar descriptores que valoren las actuaciones, es comenzar con muestras representativas de tales actuaciones. Aquí se puede preguntar a los examinadores representativos qué ven cuando trabajan con las muestras (cualitativas). Los métodos del 5 al 8 son variantes de esta idea. Como alternativa, se puede pedir a los examinadores que evalúen las muestras y después utilicen una técnica estadística adecuada para identificar las características fundamentales en las que apoyan sus decisiones (cuantitativas). Los métodos 10 y 11 son ejemplos de este enfoque. La ventaja de analizar muestras de actuaciones es que se pueden obtener descripciones muy concretas basadas en datos.

El último método, el 12, es el único que realmente *escalona* los descriptores en un sentido matemático. Éste fue el método utilizado para desarrollar los niveles comunes de referencia y los descriptores ilustrativos, después del método 2 (intuitivo) y del 8 y el 9 (cualitativos). Sin embargo, la misma técnica estadística también se puede utilizar después del desarrollo de la escala con el fin de ratificar su uso en la práctica y de identificar las necesidades de revisión.

Métodos intuitivos:

Estos métodos no requieren ninguna colección estructurada de datos, sólo la interpretación de la experiencia basada en principios.

N.º 1 **Experto:** Se pide a alguien que escriba la escala, que puede hacerlo consultando escalas existentes, documentos curriculares u otro material adecuado proveniente de fuentes, después de realizar un análisis de las necesidades del grupo meta en cuestión. A continuación, puede comprobar y revisar la escala, utilizando informadores.

N.º 2 **Comisión:** Igual al método de experto, pero atañe a un grupo menor de desarrollo y a un grupo mayor de especialistas. Los especialistas comentan los borradores y pueden trabajar intuitivamente sobre la base de su experiencia o, en su caso, sobre la base de comparaciones con los alumnos o con muestras de actuaciones. Gipps (1994) y

Scarino (1996; 1997) analizan las deficiencias de las escalas curriculares del aprendizaje de lenguas modernas en la enseñanza secundaria, elaboradas mediante comisión en el Reino Unido y en Australia.

- N.º 3** *Experiencia*: Es como el método de comisión, pero el proceso dura un tiempo considerable dentro de una institución y, en su caso, de un contexto concreto de evaluación hasta que se consigue un acuerdo por quórum. Un núcleo de personas llega a un conocimiento compartido de los niveles y de los criterios, seguido de una comprobación y retroalimentación sistemáticas con el fin de perfeccionar la redacción. Grupos de examinadores pueden analizar las actuaciones en relación con las definiciones, y las definiciones en relación con actuaciones de muestra. Esta es la forma tradicional de desarrollar las escalas de dominio de la lengua (Wilds, 1975; Ingram, 1985; Liskin-Gasparro, 1984; Lowe, 1985, 1986).

Métodos cualitativos:

Todos estos métodos suponen la participación de pequeños talleres de trabajo y grupos de informadores, y una interpretación cualitativa más que estadística de la información obtenida.

- N.º 4** *Conceptos clave: formulación*. Una vez que existe un borrador de escalas, una técnica sencilla consiste en fragmentar la escala y pedir a los informadores representativos de las personas que van a usar la escala que a) coloquen las definiciones en el orden que crean correcto, b) expliquen por qué creen que es así, y después de esclarecer la diferencia existente entre esa ordenación y la que se pretendía originalmente, y c) que identifiquen qué elementos clave les ayudaron o les confundieron. Otra mejora consiste a veces en eliminar un nivel y establecer la tarea secundaria de determinar dónde existe un vacío entre dos niveles que indique la falta de un nivel entre ambos. Las escalas de certificación de *Eurocentres* se desarrollaron de esta forma.
- N.º 5** *Conceptos clave: actuaciones*. Se comparan los descriptores con actuaciones típicas de los niveles de esas bandas para asegurar la coherencia entre lo que se describe y lo que ocurrió. Algunas guías de los exámenes de Cambridge introducen a los profesores en este proceso, comparando la redacción de las escalas con calificaciones otorgadas a ejercicios escritos concretos. Los descriptores del IELTS (*International English Language Testing System*) se elaboraron pidiendo a grupos de examinadores experimentados que identificaran muestras clave de ejercicios escritos de cada nivel y después acordaran cuáles eran las características clave de cada ejercicio. Luego, mediante debates, se determinaron las características que se consideraron representativas de distintos niveles y se incorporaron a los descriptores (Alderson, 1991; Shohamy *et al.*, 1992).
- N.º 6** *Rasgo primario*: Los informadores realizan una ordenación de las actuaciones (generalmente escritas). Después se negocia una ordenación común; para a continuación determinar y describir en cada nivel el principio según el que se han ordenado los exámenes escritos, procurando subrayar las características dominantes en un nivel concreto. Lo que se ha descrito es el rasgo (característica, constructo) que determina la ordenación (Mullis 1980). Una variante habitual es la organización en un

número determinado de conjuntos, y no en un orden de valoración. Existe también una interesante variante multidimensional del enfoque clásico; en esta versión, primero se determinan cuáles son los rasgos más importantes mediante la identificación de características clave (n.º 5 anterior) y después se clasifican las muestras de cada rasgo de forma separada. Así, al final se tiene una escala analítica y múltiple de rasgos y no una escala holística del rasgo primario.

- N.º 7** *Decisiones binarias*: Otra variante del método del rasgo primario consiste en organizar primero muestras representativas en conjuntos por niveles. A continuación, en un debate que se centra en los límites entre niveles, se determinan las características clave (como en el n.º 5 anterior). Sin embargo, la característica en cuestión se formula después como una pregunta breve de criterio con una respuesta de *sí* o *no*. De esta forma se elabora un árbol de elecciones binarias. Esto proporciona al examinador un algoritmo de las decisiones que se deben seguir (Upshur y Turner, 1995).
- N.º 8** *Valoraciones comparativas*: Los grupos debaten pares de actuaciones declarando cuál es el mejor y por qué. De esta forma, se determinan las categorías del metalenguaje utilizado por los examinadores, así como las características principales que operan en cada nivel. Estas características se pueden introducir entonces en la formulación de los descriptores (Pollitt y Murray, 1996).
- N.º 9** *Clasificación de tareas*: Una vez que existen descriptores en forma de esbozos, se puede pedir a los informadores que los organicen en conjuntos según las categorías que se supone que describen y, en su caso, según los niveles. También se puede pedir a los informadores que realicen comentarios sobre los descriptores, los corrijan o enmienden y, en su caso, los rechacen, y que determinen cuáles son especialmente claros, útiles, adecuados, etc. El banco de descriptores en que se basó el conjunto de escalas ilustrativas se desarrolló y se corrigió de esta manera (Smith y Kendall, 1963; North, 1996/2000).

Métodos cuantitativos:

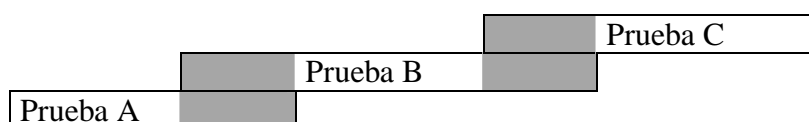
Estos métodos suponen una gran cantidad de análisis estadístico y una interpretación cuidadosa de los resultados.

- N.º 10** *Análisis discriminante*: Primero, un conjunto de muestras de actuaciones que ya han sido valoradas (preferiblemente por un equipo) se somete a un análisis detallado del discurso. Este análisis cualitativo determina y contabiliza la incidencia de distintas características cualitativas. A continuación, se utiliza la regresión múltiple para establecer cuáles de las características identificadas son importantes para determinar aparentemente la clasificación que dieron los examinadores, y esas características clave se incorporan a la hora de formular los descriptores de cada nivel (Fulcher, 1996).
- N.º 11** *Escalonamiento multidimensional*: A pesar del nombre, ésta es una técnica descriptiva que determina características clave y la relación que existe entre ellas. Las actuaciones son valoradas mediante una escala analítica de varias categorías. El producto de la técnica de análisis demuestra qué categorías fueron realmente decisivas a la hora de

determinar el nivel, y ofrece un diagrama que traza la proximidad o la distancia de las distintas categorías entre sí. Es ésta, por tanto, ésta es una técnica de investigación que determina y ratifica criterios destacados (Chaloub-Deville, 1995).

N.º 12 *Teoría de la respuesta al ítem (TRI) o análisis del rasgo latente*: La TRI ofrece una familia de medidas o de modelos de elaboración de escalas de medición. El más directo y potente es el modelo de *Rasch*, que toma su nombre de George Rasch, matemático danés. La TRI es un desarrollo de la teoría de la probabilidad y se utiliza principalmente para determinar la dificultad de los ítems individuales de pruebas que hay en un banco de ítems. Para un alumno avanzado, las posibilidades de contestar una pregunta de nivel elemental son muy altas, mientras que para un alumno de nivel elemental, las posibilidades de responder a un ítem de nivel avanzado son muy bajas. Este hecho elemental se convierte en una metodología de elaboración de escalas de medición con el modelo de Rasch, que se puede utilizar para graduar ítems de la misma escala. Un desarrollo del enfoque permite que se use para escalonar descriptores de dominio comunicativo de la lengua, así como ítems de pruebas.

En el análisis de Rasch, se puede plantear un encadenamiento de diferentes pruebas o exámenes mediante el empleo de ítems de anclaje, que son comunes a las partes adyacentes. En el siguiente diagrama, los ítems de anclaje aparecen de color gris. De esta manera, partes de la prueba se pueden diseñar para grupos concretos de alumnos, pero pueden enlazarse en una escala común. Sin embargo, hay que tener cuidado con este proceso, pues el modelo distorsiona los resultados de las puntuaciones más altas y de las puntuaciones más bajas de cada parte de la prueba.



La ventaja de un análisis de Rasch es que puede proporcionar una medición independiente de muestras y de escalas, es decir, un escalonamiento al margen de las muestras o de las pruebas y exámenes utilizados en el análisis. Se ofrecen los valores de las escalas que permanecen constantes para futuros grupos, siempre que esas futuras materias se puedan considerar grupos nuevos dentro de la misma población estadística. Los cambios sistemáticos de valores a lo largo del tiempo (por ejemplo, debido a un cambio del currículo o a la formación del examinador) se pueden cuantificar y adaptar. También se pueden cuantificar y adaptar las variaciones sistemáticas entre tipos de alumnos o examinadores (Wright y Masters, 1982; Lincare, 1989).

El análisis de Rasch se puede emplear de varias formas para valorar descriptores por escalas:

- a) Los datos de las técnicas cualitativas números 6, 7 u 8 se pueden disponer en escalas aritméticas con el análisis de Rasch.

- b) Se pueden elaborar pruebas cuidadosamente para hacer operativos descriptores de dominio de la lengua en ítems concretos de pruebas. Luego esos ítems de pruebas se pueden escalonar con el análisis de Rasch y se pueden tomar sus valores en la escala para indicar la relativa dificultad de los descriptores (Brown *et al.*, 1992; Carroll, 1993; Masters, 1994; Kirsch, 1995; Kirsch y Mosenthal, 1995).
- c) Se pueden utilizar los descriptores como ítems de pruebas para la evaluación que el profesor realiza a sus alumnos ("¿Sabe hacer X?"). Así se pueden graduar los descriptores directamente en una escala aritmética, de la misma forma en que se escalonan los ítems de pruebas de los bancos de ítems.
- d) Las escalas de descriptores incluidas en los capítulos 3, 4 y 5 se desarrollaron de esta forma. En los tres proyectos descritos en los anejos B, C y D, se ha utilizado la metodología de Rasch para escalonar los descriptores y para equiparar entre sí las escalas de descriptores resultantes.

Además de su utilidad en el desarrollo de una escala, el método de Rasch también se puede usar para analizar las formas en que se utilizan realmente las bandas de una escala de evaluación. Esto puede contribuir a resaltar la redacción deficiente, la utilización errónea o el exceso de uso de una banda, así como a perfeccionar la revisión (Davidson, 1992; Milanovic *et al.*, 1996; Stansfield y Kenyon, 1996; Tyndall y Kenyon, 1996).

Los usuarios del Marco de referencia pueden tener presente y, en su caso, determinar:

- *En qué medida se da a las calificaciones otorgadas en su sistema un sentido compartido mediante definiciones comunes.*
- *Cuál de los métodos presentados anteriormente, o qué otros métodos, se utilizan para desarrollar dichas definiciones.*

Bibliografía seleccionada y comentada: elaboración de escalas de medición del dominio de la lengua

- ALDERSON, J. C. (1991), «Bands and scores», en ALDERSON, J. C. y NORTH, B. (eds.), *Language testing in the 1990s. Developments in ELT*, London, British Council/Macmillan, págs. 71-86. Analiza los problemas causados por la confusión de finalidad y orientación, y el desarrollo de las escalas de expresión oral de IELTS.
- BRINDLEY, G. (1991), «Defining language ability: the criteria for criteria» en ANIVAN, S. (ed.), *Current Developments in language testing*, Singapur, Regional Language Centre. Crítica fundamentada de la reivindicación de las escalas de dominio de la lengua como representación de la evaluación referida al criterio.
- BRINDLEY, G. (1998), «Outcomes-based assessment and reporting in language learning programmes, a review of the issues», en *Language Testing*, 15 (1), págs. 45-85. Critica el énfasis en los resultados en función de lo que los alumnos saben hacer y no centrándose en aspectos de la competencia emergente.
- BROWN, ANNIE, ELDER, CATHIE, LUMLEY, TOM, MCNAMARA, TIM y MCQUEEN, J. (1992), «Mapping abilities and skill levels using Rasch techniques», artículo presentado en el decimocuarto *Language Testing Research Colloquium*, Vancouver; reimpresso en *Melbourne, Papers in Applied Linguistics*, 1/1, págs. 37-69. El uso clásico del escalonamiento de ítems de pruebas con el modelo de Rasch para producir una escala de dominio de la lengua, partiendo de tareas de comprensión escrita probadas en los distintos ítems.
- CARROLL, J. B. (1993), «Test theory and behavioural scaling of test performance», en FREDERIKSEN, N., MISLEVY, R. J. y BEJAR, I. I. (eds.), *Test theory for a new generation of tests*, Hillsdale, Nueva Jersey. Lawrence Erlbaum Associates, págs. 297-323. Artículo innovador que recomienda el uso del modelo de Rasch para escalonar los ítems de pruebas y elaborar así una escala de dominio de la lengua.
- CHALOUB-DEVILLE, M. (1995), «Deriving oral assessment scales across different tests and rater groups», en *Language Testing*, 12 (1), págs. 16-33. Estudio que revela los criterios que utilizan los hablantes nativos de árabe cuando valoran a los alumnos. Es prácticamente la única aplicación de un escalonamiento multidimensional de la evaluación en lenguas.

DAVIDSON, F. (1992), «Statistical support for training in ESL composition rating», en HAMP-LYONS (ed.), *Assessing second language writing in academic contexts*, Norwood, N. J. Ablex, págs. 155-166.

Una explicación muy clara de la forma de ratificar una escala de evaluación en un proceso cíclico con el análisis de Rasch. Argumenta a favor de un enfoque semántico del escalonamiento, más que por un enfoque concreto adoptado, por ejemplo, en los descriptores ilustrativos.

FULCHER (1996), «Does thick description lead to smart tests? A data-based approach to rating scale construction», en *Language Testing*, 13 (2), págs. 208-38.

Enfoque sistemático de los descriptores y del desarrollo de escalas de medición, partiendo de un análisis apropiado de lo que ocurre realmente en la actuación. Es un método que requiere mucho tiempo.

GIPPS, C. (1994), *Beyond testing*, Londres, Falmer Press.

A favor de la evaluación centrada en un nivel de exigencia realizada por el profesor, con relación a puntos comunes de referencia establecidos por medio de trabajo en grupo. Análisis de problemas ocasionados por descriptores imprecisos en el Currículo Nacional Inglés.

KIRSCH, I. S. (1995), «Literacy performance on three scales: definitions and results», en *Literacy, economy and society: Results of the first international literacy survey*, París, Organisation for Economic Cooperation and Development (OECD), págs. 27-53.

Informe sencillo y asequible sobre un uso sofisticado del modelo de Rasch para elaborar una escala de niveles partiendo de datos de pruebas. Método desarrollado para predecir y explicar la dificultad de nuevos ítems de pruebas partiendo de las tareas y de las competencias implicadas, es decir, con relación a un marco.

KIRSCH, I. S. y MOSENTHAL, P. B. (1995), «Interpreting the IEA reading literacy scales», en BINKLEY, M., RUST, K y WINGLEE, M. (eds.), *Methodological issues in comparative educational studies: The case of the IEA reading literacy study*, Washington D. C., US Department of Education. National Center for Education Statistics, págs. 135-192.

Versión más detallada y técnica que la anterior y que traza el desarrollo del método a través de tres proyectos relacionados.

LINACRE, J. M. (1989), *Multi-faceted Measurement*, Chicago, MESA Press.

Avance innovador en estadística que permite tener en cuenta el rigor de los examinadores a la hora de informar de los resultados de una evaluación. Aplicado al proyecto para desarrollar los descriptores ilustrativos con el fin de comprobar la relación de los niveles con los cursos escolares.

LISKIN-GASPARRO, J. E. (1984), «The ACTFL proficiency guidelines: Gateway to testing and curriculum», en *Foreign Language Annals*, 17/5, págs. 475-489.

Explicación de los fines y desarrollos de la escala americana ACTFL, partiendo de su escala originaria del *Foreign Service Institute* (FSI).

LOWE, P. (1985), «The ILR proficiency scale as a synthesising research principle: the view from the mountain», en JAMES, C. J. (ed), *Foreign Language Proficiency in the Classroom and Beyond*, Lincolnwood, National Textbook Company (III).

Descripción detallada del desarrollo de la escala de la *Interagency Language Roundtable* (ILR) de Estados Unidos, que procede del FSI. Funciones de la escala.

LOWE, P. (1986), «Proficiency: panacea, framework, process? A Reply to Kramsch, Schulz, and particularly, to Bachman and Savignon», en *Modern Language Journal*, 70/4, págs. 391-397.

Defensa de un sistema que funcionó bien — en un contexto específico— a pesar de la crítica académica que suscitó la difusión de esta escala y su metodología basada en entrevistas para la educación (con ACTFL)

MASTERS G. (1994), «Profiles and assessment», en *Curriculum Perspectives*, 14,1, págs. 48-52.

Breve informe de la forma en que se ha utilizado el modelo de Rasch para clasificar por escalas los resultados de pruebas y las evaluaciones que realiza el profesor, con el fin de crear un sistema curricular descriptivo en Australia.

MILANOVIC, M., SAVILLE, N., POLLITT, A. y COOK, A. (1996), «Developing rating scales for CASE: Theoretical concerns and analyses», en CUMMING, A. y BERWICK, R. *Validation in language testing*, Clevedon, Avon, Multimedia Matters, págs. 15-38.

Explicación clásica del uso del modelo de Rasch para perfeccionar una escala de medición utilizada con una prueba de expresión oral, reduciendo los niveles de la escala a un número que los examinadores puedan utilizar con eficacia.

- MULLIS, I. V. S. (1981), *Using the primary trait system for evaluating writing*, Manuscrito n.º 10-W-51, Princeton Nueva Jersey, Educational Testing Service.
- NORTH, B. (1993), «The development of descriptors on scales of proficiency: perspectives, problems, and a possible methodology», artículo en *NFLC: National Foreign Language Center*, Washington D. C., abril de 1993.
- NORTH, B. (1994), «Scales of language proficiency: a survey of some existing systems», en *CC-LANG* (94) 24, Estrasburgo, Consejo de Europa.
- NORTH, B. (1996/2000), *The development of a common framework scale of language proficiency*, Tesis doctoral, Thames Valley University. Reimpresión en 2000, Nueva York, Peter Lang.
- NORTH, B. (en prensa), Scales for rating language performance in language tests: descriptive models, formulation styles and presentation formats, artículo de investigación de TOEFL, Educational Testing Service, Nueva Jersey, Princeton.
- NORTH, B. y SCHNEIDER, G. (1998), «Scaling descriptors for language proficiency scales», en *Language Testing*, 15/2, págs. 217- 262.
- Explicación clásica de la metodología del rasgo primario en la escritura en lengua materna para desarrollar una escala de valoración.
- Crítica del contenido y del desarrollo metodológico de las escalas tradicionales de dominio de la lengua. Propuesta de un proyecto para desarrollar los descriptores ilustrativos con profesores y escalonarlos con el modelo de Rasch, partiendo de la evaluación que realiza el profesor.
- Estudio integrador de escalas curriculares y de escalas de valoración, analizadas y estudiadas posteriormente como punto de partida del proyecto para desarrollar descriptores ilustrativos.
- Análisis de las escalas de dominio de la lengua, de cómo se relacionan los modelos de competencia y el uso de la lengua con las escalas. Explicación detallada de las fases de desarrollo del proyecto de elaboración de los descriptores ilustrativos, de los problemas abordados y de las soluciones adoptadas.
- Análisis detallado y estudio histórico de los tipos de escalas de valoración, utilizadas con pruebas de expresión oral y de expresión escrita: ventajas, inconvenientes, escollos, etc.
- Visión general del proyecto de elaboración de los descriptores ilustrativos. Estudia los resultados y la estabilidad de la escala. En un anejo se muestran ejemplos de instrumentos y de productos.

- POLLITT, A. y MURRAY, N. L. (1996), «What raters really pay attention to», en MILANOVIC, M. y SAVILLE, N. (eds.) (1996), *Performance testing, cognition and assessment*, Estudios sobre evaluación lingüística 3, artículos seleccionados del decimoquinto Coloquio del *Language Testing Research*, celebrado en Cambridge y Atenas el 4 de agosto de 1993, Cambridge, University of Cambridge Local Examinations Syndicate, págs. 74-91.
- Interesante artículo metodológico que vincula el análisis de cuadros de repertorio con una técnica sencilla de escalonamiento, para determinar en qué se centran los evaluadores en los distintos niveles de dominio de la lengua.
- SCARINO, A. (1996), «Issues in planning, describing and monitoring long-term progress in language learning», en las actas del *X Congreso de Lenguas Nacionales de AFMLTA*, págs. 67-75.
- Critica el uso de la redacción imprecisa y la falta de información sobre la calidad de la actuación de los alumnos en las especificaciones descriptivas características del currículo en el Reino Unido y en Australia.
- SCARINO, A. (1997), «Analysing the language of frameworks of outcomes for foreign language learning», en las actas del *XI Congreso de Lenguas Nacionales de AFMLTA*, págs. 141-258.
- De contenido y orientación semejante al anterior.
- SCHNEIDER, G. y NORTH, B. (1999), «“In anderen Sprachen kann ich” ... Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit», en *NFP 33/SKBF* (Umetzungsbericht), Berna / Aarau.
- Breve informe sobre el proyecto de elaboración de las escalas ilustrativas. También presenta la versión suiza del Portfolio (cuarenta páginas A5).
- SCHNEIDER, G. y NORTH, B. (2000), «“Dans d’autres langues, je suis capable de ...”, Echelles pour la description, l’évaluation et l’auto-évaluation des compétences en langues étrangères», en *PNR 33/CSRE* (rapport de valorisation), Berna/Aarau.
- Semejante al anterior.
- SCHNEIDER, G. y NORTH, B. (2000), «Fremdsprachen können -was heisst das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit», en Chur/Zürich, Verlag Rüegger AG.
- Informe completo del proyecto de elaboración de las escalas ilustrativas. Capítulo sencillo sobre el escalonamiento en inglés. También introduce la versión suiza del Portfolio.

SKEHAN, P. (1984), «Issues in the testing of English for specific purposes», en *Language Testing*, 1/2, págs. 202-220.

Critica las referencias a la norma y la redacción relacionada con ello de las escalas de ELTS.

SHOHAMY, E., GORDON, C. M. y KRAEMER, R. (1992), «The effect of raters' background and training on the reliability of direct writing tests», en *Modern Language Journal*, 76, págs. 27-33.

Explicación sencilla del método básico y cualitativo para el desarrollo de una escala analítica de la expresión escrita. Proporcionó una asombrosa fiabilidad entre evaluadores que no estaban entrenados y que no eran profesionales.

SMITH, P. C. y KENDALL, J. M. (1963), «Retranslation of expectations: an approach to The construction of unambiguous anchors for rating scales», en *Journal of Applied Psychology*, 47/2.

El primer enfoque de los descriptores de valoración y no sólo de las escalas de la expresión escrita. Innovador. Su lectura es compleja.

STANSFIELD, C. W. y KENYON, D. M. (1996), «Comparing the scaling of speaking tasks by language teachers and the ACTFL guidelines», en CUMMING, A. y BERWICK, R., *Validation in language testing*, Clevedon, Avon, Multimedia Matters, págs. 124-153.

Utilización del modelo de escalonamiento de Rasch, para confirmar la ordenación de tareas que aparecen en las directrices de ACTFL. Interesante estudio metodológico que perfeccionó el enfoque adoptado en el proyecto de elaboración de los descriptores ilustrativos.

TAKALA, S. y KAFTANDJIEVA, F. (de próxima publicación), «Council of Europe scales of language proficiency: A validation study», en ALDERSON, J. C. (ed.), *Case studies of the use of the Common European Framework*, Consejo de Europa.

Informe sobre la utilización de un desarrollo más profundo del modelo de Rasch para escalonar las autoevaluaciones de tipo lingüístico en relación con adaptaciones de los descriptores ilustrativos. Contexto: proyecto DIALANG: pruebas para el finlandés.

TYNDALL, B. y KENYON, D. (1996), «Validation of a new holistic rating scale using Rasch multifaceted analysis», en CUMMING, A. y BERWICK, R., *Validation in language testing*, Clevedon, Avon, Multimedia Matters, págs. 9-57.

Explicación sencilla de la validación de una escala para la valoración de entrevistas en inglés como segunda lengua para el examen de ingreso en la universidad. Uso clásico del modelo multidimensional de Rasch para determinar las necesidades de entrenamiento.

UPSHUR, J. y TURNER, C. (1995), «Constructing rating scales for second language tests», en *English Language Teaching Journal*, 49 (1), págs. 3-12.

Desarrollo sofisticado de la técnica del rasgo primario para elaborar cuadros de decisiones binarias. Muy adecuado para el sector de educación escolar.

WILDS, C.P. (1975), «The oral interview test», en SPOLSKY, B. y JONES, R. (eds.), *Testing language proficiency*, Center for Applied linguistics, Washington D. C., págs. 29-44.

La primera aparición de la escala original de valoración del dominio de la lengua. Vale la pena leerlo atentamente para ver matices que se han perdido en la mayoría de los enfoques de entrevista desde entonces.