Diseño y validación de exámenes de dominio de lengua

Joaquín M. Cruz Trapero

Instantánea, Biarritz, 1906: Joaquín Sorolla y Bastida.
Fuente: Fundación Museo Sorolla.

Diseño y validación de exámenes de dominio de lengua

Joaquín M. Cruz Trapero

MINISTERIO
DE EDUCACIÓN, FORMACIÓN PROFESIONAL
Y DEPORTES

Catálogo de publicaciones del MEFD: https://www.libreria.educacion.gob.es/
Catálogo general de publicaciones oficiales: https://cpage.mpr.gob.es/

Título de la obra:
Diseño y validación de exámenes de dominio de lengua

Autor: Joaquín M. Cruz Trapero

MINISTERIO DE EDUCACIÓN,
FORMACIÓN PROFESIONAL
Y DEPORTES

Dirección General de Evaluación
y Cooperación Territorial

Edita:

Subdirección General de Atención
al Ciudadano, Documentación y
Publicaciones

Edición: 2024
NIPO en papel: 164-24-225-9
NIPO en línea: 164-24-226-4
Depósito legal: M-24952-2024

Maquetación: Imprenta Gamar, S.L.

Prólogo

Este libro acerca al lector hispanohablante las experiencias que han transformado la evaluación de lenguas durante los últimos treinta años. En este periodo, la evaluación de lenguas ha abandonado el rincón de la lingüística aplicada anglófona en el que se encontraba recluida y se ha convertido en una profesión internacional dotada de un metalenguaje común y estándares transparentes.

Aunque se conocen ejemplos previos de exámenes (principalmente el *Cambridge Certificate of Proficiency* de 1913), tres hitos desencadenan el nacimiento la evaluación de lenguas como disciplina académica a mediados del siglo XX: el libro *Language Testing*, publicado en 1961 por Robert Lado, de la Universidad de Georgetown; la aparición del TOEFL (*Test of English as a Foreign Language*) en 1964; y el primer *Language Testing Symposium* celebrado en la Universidad de Edimburgo en 1968. Las generaciones de profesionales norteamericanos posteriores a estos hechos, influidas por la psicología aplicada, enfatizaron la medición en detrimento de la creación de tareas orientadas a la vida real. Mientras tanto, en gran parte de Europa (incluyendo Gran Bretaña) la práctica se centró en construir tareas desarrolladas y revisadas por expertos en detrimento del análisis psicométrico.

Esta situación comenzó a cambiar en 1990, cuando un estudio comparativo entre los exámenes de Cambridge y TOEFL, y la fundación ese mismo año de ALTE (Association of Language Testers in Europe) hicieron que la tradición estadounidense y la europea convergieran. Todo ello tuvo su reflejo en la investigación que desembocó en la publicación del *Marco Común Europeo de Referencia* (MCER) (Consejo de Europa, 2001) y en la forma en que muchos exámenes europeos fueron completamente reformados o desarrollados desde cero tras dicha publicación. Desde entonces, la aparición del manual *Relating Language Examinations to the Common European Framework of Reference for Languages* (Consejo de Europa, 2009) y la fundación en 2004 de EALTA (European Association for Language Testing and Assessment) han propiciado en Europa un incremento

constante de la formación en evaluación de lenguas y han dado lugar a exámenes más equilibrados, válidos y fiables. El presente libro contribuye de forma significativa a extender este proceso en el mundo hispanohablante, sobre todo si se tiene en cuenta la escasa formación sobre evaluación de lenguas que hasta la fecha se ha venido ofreciendo en los departamentos universitarios.

Como el autor explica en su introducción, esta obra es, en realidad, dos libros: un texto que es un placer leer, que aporta muchas ideas, y un manual de referencia al que acudir mientras se prepara un examen de dominio de lengua. El Capítulo 1 reflexiona sobre qué es el lenguaje, sobre la forma en que este se ha estudiado a lo largo de los siglos (con la primera *gramática* de una lengua vulgar escrita en castellano) y sobre cómo los avances en neurolingüística vinculados a la inteligencia artificial pueden llevar a mecanizar la evaluación de dominio de lenguas en el marco de la biolingüística.

El Capítulo 2 ofrece un resumen conciso de la psicometría clásica (índices de dificultad y discriminación, fiabilidad, etc.) y de la más reciente Teoría de Respuesta al Ítem (TRI), que es utilizada principalmente para desarrollar bancos de ítems para diferentes niveles y la base sobre la que se calibraron los descriptores de los niveles descritos en el MCER. Como señala el autor, con nuestras pruebas y exámenes tan solo podemos trabajar con el «reflejo» de la competencia lingüística de los candidatos, que no es directamente observable y cuya evaluación justificamos a través de la psicometría.

Tras recorrer la historia de las pruebas de dominio de lengua y después de considerar los desafíos que la actual industrialización de estas puede suponer para su validez, el Capítulo 3 explica la diferencia entre medición, evaluación y exámenes e introduce el *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación* (Consejo de Europa, 2001) en el contexto de la integración de Europa. Es en este punto en el que el libro se enfoca en su objetivo principal. Yendo de lo teórico y general a lo concreto y práctico, ofrece al lector un resumen de los diversos aspectos que componen un argumento de validez, que quedan ilustrados en una útil tabla. A continuación, el autor habla de cómo se han de definir constructos y especificaciones y de cómo seleccionar tareas para evaluar los diferentes modos de comunicación, abordando incluso el reto que supone examinar las destrezas de forma integrada y el desafío que implica evaluar la mediación. Cada sección de este capítulo comienza con un análisis de la naturaleza de la actividad en la que se centra; por ejemplo, en el análisis de la comprensión auditiva se analizan detalladamente las condiciones físicas necesarias y la naturaleza de los textos utilizados, al tiempo que se ofrecen consejos para desarrollar diferentes tipos de tareas, algo que será particularmente útil si tenemos en cuenta que las pruebas de comprensión auditiva son las más difíciles de preparar. En esta sección el lector también encontrará una lista muy útil de puntos que se deben tener en cuenta al utilizar ítems de respuesta múltiple.

Finalmente, el Capítulo 4 está dedicado al vital y complejo tema de diseñar y validar (cualitativa y cuantitativamente) escalas de evaluación para medir las habilidades productivas, un área muy descuidada en las pruebas de lenguas.

Es raro que una obra sobre evaluación de lenguas cubra todos los temas cualitativos y cuantitativos relevantes en apenas 200 páginas. Este libro es una guía de referencia muy clara y útil para todos aquellos que necesiten diseñar o revisar exámenes de dominio de lengua.

Brian North
Coautor del MCER, 2001 y del MCERVC, 2020

ÍNDICE

Pág.

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Introducción

Cuando defendí mi tesis doctoral, centrada en el diseño de escalas analíticas, uno de los miembros del tribunal calificó mi definición del lenguaje de «impresionista». Utilizó la palabra «impresionista» no porque lo que escribí le hubiese recordado a un cuadro de Sorolla, sino más bien para criticar el hecho de que mi definición del lenguaje estaba basada, principalmente, en opiniones e impresiones personales. En parte, era cierto.

Por aquel entonces, en 2017, acababa de comenzar a leer sobre la biolingüística y estaba bajo el embrujo de una deriva de investigación que proponía, entre otras cosas, partir de la biología para el estudio del lenguaje humano. Ni la biolingüística se reduce a esa idea, ni la idea en sí era nueva, pero para mí, en aquel momento, los biolingüistas se dibujaban como unos aguerridos investigadores que, en muchos casos, llegados de territorios científicos lejanos, habían decidido apartarse de su zona de confort para explorar nuevos campos del conocimiento que pudieran llevarles a entender qué es el lenguaje. Esta última fue precisamente una de las preguntas que el antes mencionado miembro del tribunal me planteó: «¿qué es para usted el lenguaje humano?». Tras la pregunta me sentí desnudo y sin respuesta y, de hecho, durante tiempo tuve la sensación de que aquellas páginas y el tiempo que les había dedicado no alcanzaban para dar respuesta a una cuestión que debería estar en el corazón de todo lingüista. «¿Qué es el lenguaje humano?» Con los años he comprendido que mi trabajo no fue inútil y que, sencillamente, un fenómeno tan complejo como el lenguaje no puede definirse con aforismos ni contenerse en un resumen, por brillante que este sea.

En mi búsqueda de respuestas para aquella pregunta he llegado a la conclusión, nada original, de que para definir algo o para estructurar teorías sobre ello es fundamental aprender a medir. Durante todo el tiempo que los químicos han estado midiendo el peso atómico del antimonio, en las ciencias sociales nos hemos contentado con aproximaciones que en muchos casos no son operativas ni

manejables. ¿Por qué a los lingüistas nos cuesta tanto medir las palabras más allá de los espacios en blanco que separan unas de otras?

Desafortunadamente, el presente libro sigue sin dar una definición clara de lo que es el lenguaje. Sin embargo, a pesar de esta carencia, creo sinceramente que el libro es relevante por lo que aporta a la forma en que se pueden diseñar exámenes de dominio de lengua.

Fundamentalmente, esta obra gira en torno a la definición del lenguaje, su medición y cómo ambos aspectos se combinan en los exámenes. Para tratar estos tres temas he dividido el libro en cuatro capítulos, de los cuales los dos primeros son teóricos y los dos segundos prácticos. Tal y como Brian North explica en el prólogo, el capítulo 1 es una breve reflexión sobre el lenguaje que parte de las concepciones más antiguas del mismo y se extiende hasta las más modernas. En el capítulo 2, se expone un resumen de conceptos psicométricos básicos y se identifican algunas de las herramientas que se pueden utilizar para medir el lenguaje. Los capítulos 3 y 4 están dedicados al diseño de los exámenes. Dentro de este bloque, el capítulo 3 es el que contiene la explicación práctica de cómo debe idearse y redactarse una prueba de dominio de lenguas, mientras que el capítulo 4 profundiza en el diseño y validación de escalas de evaluación.

A su manera, este libro es, igual que la maravillosa *Rayuela* de Cortázar ([1963] 2016), muchos libros pero, sobre todo, es dos libros. El primero, el que describo en el párrafo anterior, se deja leer en la forma corriente y está destinado al lector que dispone de tiempo suficiente. Al terminar el libro, este lector no apresurado tendrá la sensación de haber reflexionado sobre el lenguaje, la medición y sobre los exámenes de dominio de lengua. Otro tipo de lector, el que solo desee leer acerca de la confección de exámenes partiendo desde cero, puede comenzar leyendo la sección 3.4, pasar a la 3.5 y acabar en la 4.2, renunciando sin remordimiento al resto de páginas.

Este libro me ha acompañado durante varios años, prácticamente desde aquel día de 2017 en el que me hicieron la referida pregunta sobre el lenguaje. Las ideas aquí contenidas, como escribía Machado, se han lanzado en busca de palabras. Espero, querido lector, que el tiempo que invierta en su lectura sea productivo y que este vademécum le acompañe en su trabajo y en sus reflexiones como una herramienta útil.

Joaquín M. Cruz Trapero
Autor del libro

CAPÍTULO 1

El Lenguaje

Garabatea un círculo y aparecerá π. Adéntrate en un nuevo sistema solar y encontrarás las fórmulas Tycho Brahe acechando en el negro y aterciopelado manto del espacio/tiempo. Pero ¿dónde esconde el universo palabra alguna bajo su capa externa de biología, geometría y roca insensible? […] Cualquier tipo de vida inteligente con la que nos hemos topado (los globos de Jove II, los Constructores de Laberintos, los émpatas Seneschai de Hebrón, la Gente Palo de Durulis, los arquitectos de las Tumbas del Tiempo o el propio Alcaudón) nos han legado misterios y oscuros artefactos pero ni rastro de lengua alguna. Ni una sola palabra. (Simmons, 2004:139)

De Saussure (1994:18) apuntó que si una ciencia no es capaz de determinar la naturaleza de su objeto de estudio tampoco es capaz de procurarse un método. Apenas treinta años después de que se publicase la obra de de Saussure, Skinner (1957:5) apuntó que la creación de una verdadera ciencia del comportamiento verbal humano había estado históricamente lastrada por motivos espurios (*fictitious* es el término que usa Skinner, *idées absurdes*, *préjugés*, *mirages* y *fictions* los que usa de Saussure). En efecto, el estudio del lenguaje humano ha estado lastrado por mitos, convenciones o tabúes como la propia crítica de Skinner al uso de la estadística (Rasch, 1980:xx), o la famosa prohibición de la Societé de Linguistique de Paris, que en sus primeros estatutos dejaba claro que no admitiría ninguna comunicación concerniente al origen del lenguaje (Arellano, 1979:9) por miedo a que esto suscitase debates estériles o acalorados. Actitudes como estas nos han dificultado alcanzar definiciones centrales para nuestra disciplina y nos han impedido dotarnos de las mismas herramientas con que se cuenta en otros campos del saber. Si el astrónomo es capaz de determinar con precisión la posición de una estrella, ¿por qué los lingüistas no hemos podido consensuar una definición exacta del lenguaje a pesar de habernos interesado por ello desde épocas remotas?

La respuesta a esta pregunta se haya, en parte, en la forma en que se han estudiado el «lenguaje» (la «[f]acultad del ser humano de expresarse y comunicarse con los demás a través del sonido articulado o de otros sistemas de signos» (RAE, 2024)), la «lengua» (el «[s]istema de comunicación verbal propio de una comunidad humana que cuenta generalmente con escritura» (RAE, 2024)) y la «lingüística» (la «[c]iencia del lenguaje» (RAE, 2024)). Como veremos en el breve recorrido histórico que sigue, siglos de estudio no nos han bastado para establecer los límites de nuestra ciencia porque hemos obviado el objeto del que esta emana para centrarnos en el estudio de la lengua, el reflejo del lenguaje en las paredes de nuestra caverna.

En este recorrido histórico, los textos de los grandes historiadores de la lingüística (Thomsen, 1945; Robins, 1967; Malmberg, 1974; Arellano, 1979 o Mounin, 1995) evidencian que el interés por la capacidad humana del habla no estalla

como una tempestad en el cielo sereno de nuestra época, sino que se desarrolla de forma intuitiva en distintas y distantes culturas. En Oriente, por ejemplo, las primeras reflexiones sobre el lenguaje surgieron hace veinticuatro siglos. En Occidente, según de Saussure (1994:15), «[l]a ciencia que se ha constituido en torno a los hechos de la lengua ha pasado por tres fases sucesivas antes de reconocer cuál es su verdadero y único objeto». La primera de estas fases, que para de Saussure surge en la Grecia clásica, sería la que comenzó por organizar la gramática. En la segunda fase, que data del siglo XVIII, asistimos al nacimiento de la filología con los trabajos de Wolf. Finalmente, el tercer periodo, que comenzaría en el siglo XIX, es el correspondiente a la filología comparativa, dentro del cual se enmarca el nacimiento de la lingüística propiamente dicha. Es evidente que existe una cuarta fase, eminentemente empírica y global, que comienza con el propio de Saussure quien, bien por humildad, bien por falta de perspectiva histórica, no atribuye a su propio trabajo la importancia que este realmente tiene.

Como decíamos, las primeras reflexiones sobre el lenguaje que se preservan provienen de Oriente y están vinculadas a los Vedas hindúes. Dado que era de suma importancia que estos cantos sagrados se transmitiesen con la mayor exactitud posible, ya que de ello derivaba su valor religioso (Thomsen, 1945:14), su interpretación se acompañó de los vedangas, textos sobre las distintas ramas de la ciencia védica entre las que se incluyen la fonética, la métrica, la gramática, la etimología y el léxico. En concreto, los vedangas escritos por Panini (cuyo origen se remonta a los siglos IV, V o VI a. C., según las fuentes) son los más relevantes en este ámbito y, con razón, la admiración de los filólogos modernos, pues llegaron hace unos veinticuatro siglos a una perfección en el análisis lingüístico solo igualado por otros gramáticos en tiempos modernos (*vid.* Arellano, 1979:21). Sin abandonar Oriente, parece lógico pensar que los chinos, que atesoran la tradición escrita ininterrumpida más larga de la historia, también reflexionaran sobre el lenguaje. De hecho, según nos cuenta Wang (1989:183), prácticamente al mismo tiempo que Grecia debatía acerca de la naturaleza de las palabras, el filósofo chino Xunzi escribía, en el siglo III a. C., sobre temas parecidos. El egipcio, un pueblo que poseía un notable conocimiento de ciencias tan complejas como la astronomía, las matemáticas o la arquitectura, a buen seguro también reflexionó sobre estos temas en algún punto remoto de su vasta historia. Aunque se conocen pocos trabajos egipcios centrados en la lingüística, restos arqueológicos como la piedra Rosetta, escrita en egipcio, demótico y griego demuestran que la reflexión lingüística debió de existir sin duda en Egipto.

En la Grecia clásica (la primera de las fases del estudio del lenguaje que identifica de Saussure) siglos de guerras e invasiones modificaron profundamente las costumbres de los pueblos egeos. Es fácil imaginar que en esta infinidad de intercambios lingüísticos, potenciados por la guerra y el comercio, los hablantes de distintos dialectos (aqueo, jónico, ático, etc.) llegasen a razonar sobre las for-

mas más prácticas de expresarse con un grado de sofisticación tal que los llevase a una primera e intuitiva visión de la lingüística. En el siglo V a. C., Platón (2004) trufa de reflexiones lingüísticas su diálogo *Crátilo*. Heródoto, coetáneo de Platón, narra a caballo entre la historia y el mito, en el segundo libro de sus *Historias*, el extraño experimento lingüístico del faraón egipcio Psamético (Heródoto, 1992:278–280) para averiguar cuál era el pueblo más antiguo de la Tierra. Aristóteles impulsó el estudio de las categorías gramaticales y de los casos en la lengua griega (Thomsen, 1945:24) que, de alguna manera, perviven hasta nuestros días. A través de Grecia, Occidente también hereda la escritura de los fenicios, a los que debemos la democratizaron la escritura, que pronto trascendió lo contable y se extendió a otras funciones, incluida la metalingüística.

Roma tomaría mucho de Grecia y, al igual que esta, se interesó más por la lengua que por el lenguaje. Aunque la mayor parte del estudio lingüístico en Roma estuvo basado en la herencia griega, hoy consideramos a los latinos, y no a los griegos, como el fundamento de nuestra enseñanza tradicional en lo relativo a la gramática (Mounin, 1995:100). Tras la caída del imperio romano en el siglo V, el latín continuó siendo la lengua franca de la cultura y gran parte de la tradición lingüística greco-latina pasó a la Edad Media (Robins, 1967:69) a través de la iglesia católica, en el seno de la cual se desarrollan trabajos tan importantes como las *Etimologías* de San Isidoro de Sevilla (1951), del siglo VII. Al contrario de lo que este título sugiere al lector contemporáneo, las *Etimologías* (*ibid*.) no se centran en cuestiones de morfología o de lingüística diacrónica; son un compendio medieval del saber en el que San Isidoro habla de aspectos tan dispares como gramática, dialéctica, matemáticas (en las que incluye música y astronomía), medicina, leyes, iglesia, antropología, zoología, arquitectura o geografía. Impresiona imaginar las vicisitudes con las que San Isidoro tuvo que lidiar para tener acceso a los códices que constituyeron el saber plasmado en su obra, particularmente en un tiempo en que ni los libros eran tan comunes ni el saber era tan codiciado.

La aparición de la imprenta en 1440 nos conduce al Renacimiento. Si los fenicios democratizaron la escritura, Gutenberg democratizó la lectura y propició que un número cada vez mayor de obras se distribuyesen en todo el mundo. En esta misma época afloran los primeros estudios de lenguas romances, por ejemplo, la conocida *Gramática castellana*, de 1492, de Antonio de Nebrija (la primera del español y la primera de una lengua vulgar que se publica en Europa). A este trabajo sobre una lengua vernácula seguirían otros dedicados al italiano o al polaco en Europa, al nahual en Méjico, al quechua en Perú o al guaraní en Brasil. Sabemos que en la América Precolombina existieron al menos ciento veintitrés familias de lenguas. Al tratarse de lenguas sin tradición escrita conocida a día de hoy, es difícil determinar el nivel de reflexión lingüística que dichas lenguas alcanzaron (*vid*. Malmberg, 1974:61–104). De haber existido, dicha reflexión lingüística podría haber tenido lugar incluso al margen de la escritura. La cultura inca peruana, por

ejemplo, conquistó y gobernó un poderoso imperio sin apoyo de la escritura (más allá de un sistema de mensajes mediante nudos en cuerdas, o quipus) (Vallejo, 2020:100). Es difícil imaginar que todo esto ocurriese al margen de una mínima reflexión lingüística.

Con la llegada de la Ilustración (la segunda de las fases que identifica de Saussure) William Jones vincula el sánscrito con el latín, el griego y las lenguas germánicas (Robins, 1967:134) abriendo el sendero que habría de seguir Rask en el siglo XIX y que cristalizaría en las famosas leyes formuladas por su colega Jacob Grimm en la *Deutsche Grammatik*, publicada entre 1819 y 1837. Jacob Grimm, junto con su hermano Wilhelm, también publicaría en 1812 la recopilación de cuentos tradicionales que popularizó para siempre las historias de Blancanieves, Hansel y Gretel o Caperucita Roja.

De esta manera llegamos al siglo XX y al *Cours de linguistique g*énérale de de Saussure (1994), publicado originalmente en 1916. Este es el primer intento moderno, cumbre del positivismo lingüístico, de dotar de un sustrato científico al conjunto del estudio de la lengua. La obra de de Saussure aporta una visión lógica, ordenada, profunda y genuinamente científica de la que se seguirá hablando durante siglos. Sin lugar a dudas, uno de los logros fundamentales del pensamiento de de Saussure es el ubicar la definición de la lengua en el centro del debate lingüístico, algo que, como decíamos al comienzo del capítulo, los lingüistas nos hemos negado a nosotros mismos durante largo tiempo. También en el siglo XX llegaron Skinner, Chomsky y una pléyade de nuevas escuelas que, siguiendo la estela de de Saussure, esta vez sí se esforzaron por definir su objeto de estudio. Aunque no olvidamos los trabajos de Luria, Vygotsky, Leontiev, Boas, Sapir, Whorf, Bloomfield, Halliday y tantos otros, en nuestra opinión, el trabajo de de Saussure, el debate entre Skinner y Chomsky, y el trabajo de Lenneberg (del que hablaremos a continuación) son los hitos que hacen del siglo XX el siglo de las luces para el estudio científico del lenguaje.

Y es que en el siglo XX se ha avanzado tanto en el estudio del lenguaje como en los veinticuatro siglos que separan a Panini de de Saussure. En parte, esto ha sido así porque nos hemos atrevido a transitar avenidas alejadas de la tradición a hombros de disciplinas como la medicina, las matemáticas o la computación, desdibujando las fronteras entre unas y otras e integrando campos de estudio sin relación aparente. Al incorporar nuevos métodos y disciplinas, podría parecer que nos alejamos del propósito de definir el objeto de estudio de la lingüística. Más bien al contrario, debemos interpretar estas incorporaciones como una manera de delimitar la fuente de nuestro estudio mediante métodos antes desconocidos, como un objeto que emana del mundo natural y que, al igual que otros fenómenos (como la gravedad, la consciencia, la luz, la vida o los movimientos de sístole y diástole del corazón humano) es susceptible de ser analizado con métodos empíricos (*vid*. Chomsky, 2000:106). Por ejemplo, aplicar los axiomas de la Teoría

Integrada de la Información (existencia, composición, información, integración y exclusión) (Massimini y Tononi, 2018) a la definición del lenguaje podría arrojar luz sobre su naturaleza última.

Fruto de los muchos avances experimentados en el siglo XX surge también la biolingüística. Según los postulados de esta disciplina, dado que el lenguaje es un fenómeno específicamente humano, en su estudio se ha de partir de los fundamentos biológicos de nuestra especie. Lenneberg abrió este debate durante la segunda mitad del siglo XX con su *Biological Foundations of Language* (Lenneberg, 1967; 1975), una obra que acercó a la lingüística y a la biología y orientó la atención de ambas disciplinas hacia el organismo que posibilita el lenguaje (*vid*. Boeckx y Piattelli-Palmarini, 2005; Di Scuillo y Boeckx, 2011; Martins y Boeckx, 2016). A partir de las reflexiones iniciales de Lenneberg se acuña el término «biolingüística» en 1974 (Martins y Boeckx, 2016:2) para hacer referencia a los estudios multidisciplinares de base biológica sobre el lenguaje. Dada la variedad de visiones y disciplinas envueltas, el término «biolingüística» aún hoy significa cosas distintas para diferentes científicos. Hay 1) quienes ven el vocablo como un nuevo nombre para la lingüística teórica de corte generativista y 2) están aquellos que lo usan para designar a los estudios que, alejándose de la lingüística teórica tradicional, tienen una firme orientación biológica (*ibid*.:1), como ocurre en estas páginas.

En los estudios lingüísticos de base biológica se suele tener en cuenta tanto la filogenia (la disciplina que estudia la historia de la evolución de una especie y su relación con otras) como la ontogenia (la ciencia que estudia los acontecimientos que tienen lugar durante el desarrollo de un individuo desde su etapa embrionaria hasta convertirse en adulto). Gracias a una perspectiva evolutiva, por ejemplo, hemos aprendido que el lenguaje humano podría haber dotado a nuestra especie de determinadas ventajas sobre otros homínidos, convirtiéndose así en una herramienta crítica para nuestra evolución (*vid*. Martin y Doumas, 2017:27), lo que nos habría permitido pasar de ser un animal más en la sabana a convertirnos en la primera especie del planeta. Mediante la perspectiva de la ontogenia es posible, por ejemplo, observar los mecanismos que propician (o impiden) la adquisición del lenguaje. El estudio del gen FOXP2, que regula el desarrollo y el funcionamiento de determinados circuitos corticotalamoestriatales, es un ejemplo de ello (*vid*. Benítez, 2016).

La lingüística del futuro quizás tenga más que ver con la realidad que describen estos estudios que con la lingüística tradicional que hoy conocemos. Tal vez, en un momento dado, la lingüística acabe siendo absorbida por la biología en una deriva «naturalista», es decir, en la búsqueda de teorías sobre el lenguaje que puedan unificarse con las ciencias naturales y que conduzcan eventualmente a la unificación de la lingüística con el núcleo de estas últimas (Chomsky, 2000:106). En una primera fase, la biología y la lingüística probablemente seguirían evolucionando de forma paralela para más tarde dar lugar a una disciplina de mayor

espectro y más propensa a la experimentación, que se acabaría convirtiendo en el tronco del que emanen las disciplinas hasta ahora desarrolladas por la lingüística tradicional, que pasarían así a ser utensilios al servicio de la biolingüística.

Si bien esto entra en el terreno de la especulación, lo que ya es una realidad es la aportación de las ciencias naturales al estudio del fenómeno humano del habla. En la actualidad es posible detectar la esclerosis lateral amiotrófica a través de la pronunciación de vocales (Tena *et al.*, 2021) y se puede transcribir el pensamiento lingüístico humano mediante resonancia magnética funcional (Tang *et al.*, 2023). También empezamos a saber cómo el cerebro humano computa la información lingüística y qué lo hace sensible a determinados estímulos y no a otros (Ding *et al.* 2016), que pueden ser recreados artificialmente en una red neuronal (Martin y Doumas, 2017).

Poder reproducir artificialmente la forma en que el cerebro humano procesa el lenguaje puede suponer avances insospechados para nuestros exámenes. Pensemos, por un momento, que somos capaces de calibrar y recrear en una red neuronal los niveles de comprensión lectora y auditiva de un hablante de, digamos, nivel B1. Si hacemos que esta red neuronal procese un determinado texto o un determinado audio a través de inteligencia artificial podremos ver cómo reacciona ante dichos estímulos. Veremos, por ejemplo, que la red neuronal de nivel B1 no reacciona ante estímulos de nivel C2 porque no los entiende. Podríamos concluir que, allí donde nuestra red neuronal se activa al procesar un estímulo concreto, debería también activarse el cerebro de un candidato. Al ser capaces de reproducir en un laboratorio los procesos cognitivos de un candidato de nivel B1 podríamos generar ítems y claves más fiables. Recorriendo el camino inverso, si fuésemos capaces de medir los campos electromagnéticos del cerebro de nuestros candidatos mientras leen o escuchan, y si pudiésemos comparar estas mediciones con una clave previamente obtenida, podríamos obtener en tiempo real la calificación de estos tan solo con observar sus cerebros mediante magnetoencefalografía. Aunque pueda parecer ciencia ficción, ya existen neuroprótesis que permiten escuchar en tiempo real el pensamiento de personas cuya capacidad de habla se encuentra totalmente impedida (Card *et al.* 2024), y se han formulado modelos matemáticos que simulan las respuestas de candidatos que en el futuro podrían sustituir a los seres humanos en el pilotaje de exámenes (Štěpánek *et al.*, 2023).

Como se apreciará si se está leyendo este libro de forma corriente, hay un capítulo ausente entre el que aquí concluye y los que siguen. ¿Cómo conciliar todas estas ideas sobre una nueva forma de lingüística con el resto de capítulos, en los que nos limitamos a analizar las manifestaciones superficiales del lenguaje? Es, sencillamente, imposible. El capítulo que defina la manera en que se ha de diseñar un examen de dominio de lengua de base biolingüística lo escribirán otros autores en el futuro. Es posible que ese capítulo invalide gran parte de lo que ahora sigue. En ese capítulo las destrezas lingüísticas, los modos de comunicación, el concepto

de tarea o el análisis psicométrico de las pruebas tal vez formen parte de un pasado tan lejano como el de las disquisiciones platónicas sobre el lenguaje. En ese capítulo, tal vez los exámenes de dominio de lengua se definirán como formas de reconstruir el pensamiento lingüístico a través de técnicas de imagen por resonancia magnética funcional o mediante magnetoencefalografía, en los que será posible corregir la producción escrita u oral de candidatos mediante redes neuronales debidamente entrenadas, como ya ocurre con los distintos sistemas de inteligencia artificial con los que cada vez estamos más familiarizados. En cualquier caso, hasta que seamos capaces de ver con más claridad tendremos que contentarnos con analizar las sombras que se proyectan en las paredes de nuestra caverna, y eso es precisamente lo que hacemos en los siguientes capítulos.

CAPÍTULO 2

La Medición

Por el número, pues, nos instruimos para no ser engañados; quita el número a las cosas y todas ellas perecen; quita al tiempo el cómputo y todo queda envuelto en la ciega ignorancia, ni puede ser diferenciado el hombre de los demás animales, que desconocen la noción del cálculo. (San Isidoro de Sevilla, 1951:76)

Medir consiste en asignar a las cosas números que implican valores o propiedades, y en hacerlo mediante reglas (Stevens, 1946:677). Asignar un número a una cosa, no obstante, no siempre conlleva una medición. No medimos, por ejemplo, cuando numeramos las casas de una calle (Campbell y Jeffreys, 1938:122). Para poder hablar de medición, además, los números asignados han de estar referidos a una unidad (gramos, metros, julios, grados, etc.) que nos permita averiguar cuántas veces está dicha unidad contenida en el objeto medido. De esto se desprende que en cualquier medición serán fundamentales tanto la unidad de medida como el procedimiento que se siga para identificar las veces que esta está contenida en el objeto medido.

Tanto las unidades de medida como los procedimientos de medición pueden variar entre disciplinas. El ser humano ha demostrado gran obstinación por medir determinados atributos que *a priori* parecen ajenos a cualquier tipo de cuantificación. Hoy nos resulta casi intuitiva la medición de los ciclos lunares o de los cambios de estación y, sin embargo, el reto intelectual que esto debió de suponer para nuestros antepasados es comparable en su complejidad al desafío que actualmente supone medir la distancia entre dos estrellas. Hemos logrado cuantificar fenómenos tan esquivos como el tiempo, la música, el peso o la temperatura. Hoy podemos leer sobre las dimensiones del ángulo del arcoíris (Crosby, 1997:19) y tenemos libros de astrofísicos que cuantifican la estructura de un arte marcial (Kitaura, 2020), al tiempo que cada vez es más frecuente medir la felicidad con escalas como el Oxford Happiness Questionnaire (Hills y Argyle, 2002). Incluso existen formas de cuantificar matemáticamente la consciencia humana (Massimini y Tononi, 2018).

Una medición objetiva nos permite, además, ubicar elementos a lo largo de un continuo. Como científicos, hemos de ser capaces de observar (¿cómo están comportándose los candidatos en esta prueba?), de medir (¿qué calificación tienen todos ellos en las diferentes partes de la prueba?), de ordenar elementos (¿qué candidatos han demostrado un dominio mayor de la lengua?) e incluso, en la me-

dida de lo posible, de predecir eventos futuros (¿serían estos candidatos capaces de usar con éxito sus habilidades lingüísticas en una situación real?).

La observación de los fenómenos lingüísticos es relativamente sencilla al tiempo que limitada, ya que se ciñe a las manifestaciones superficiales del lenguaje. Cuando analizamos cómo se comporta un candidato ante una prueba de comprensión lectora, en realidad estamos realizando una observación indirecta. Solo habrá una observación directa cuando seamos capaces de analizar lo que ocurre en el cerebro mientras produce o decodifica información lingüística. Aun en este nivel superficial e indirecto de análisis pugnamos contra interrogantes destacables. ¿Son las palabras y las oraciones medibles más allá de su extensión ortográfica o de su realización fonética? Como veremos a lo largo del presente capítulo, la respuesta es sí. En las secciones que siguen describiremos cómo las palabras se pueden convertir en números interpretables de una forma objetiva a través de la psicometría.

2.1. Psicometría

Es posible que los exámenes de dominio de lengua del futuro tengan poco que ver con los que conocemos actualmente. En ausencia de una ventana al funcionamiento del cerebro humano hemos de conformarnos con analizar las competencias lingüísticas de nuestros candidatos. Estas competencias, para mayor complicación, no son directamente observables, como ya hemos indicado. Una solución parcial a este problema es utilizar tareas con las que el candidato pueda interactuar como lo haría en una situación real de uso de la lengua y analizar el producto de esta interacción. Con todo, por acertadas y realistas que sean estas tareas y los productos de ellas derivadas, siempre serán aproximaciones, reflejos fragmentados del verdadero dominio lingüístico de nuestros candidatos. Es precisamente para analizar estos reflejos para lo que necesitamos la psicometría.

Igual que los químicos pueden contar electrones y los astrónomos medir la distancia entre estrellas, la psicometría nos permite cuantificar habilidades psicológicas, atributos y características humanas (*vid*. Buchanan y Finch, 2005) tan complejas como la inteligencia o el dominio de una lengua. Manejar la psicometría es el equivalente a manejar un potentísimo microscopio mediante el que podemos escrutar la intrincada naturaleza del comportamiento lingüístico de nuestros candidatos (*vid*. McNamara y Knoch, 2012:567). Aplicada a la confección de exámenes de dominio de lengua, la psicometría nos permite convertir las palabras en números. La psicometría nos dice dónde nuestros exámenes son valles o montañas para los candidatos y dónde hay piedras que puedan hacerles tropezar.

Sorprende que una herramienta tan potente como la psicometría nos haya sido ajena a muchos lingüistas durante tan largo tiempo. Se nos ocurren varias

explicaciones para ello. En primer lugar, las titulaciones universitarias de nuestro campo están orientadas a formar a lingüistas, estudiosos de la literatura o profesores de lengua. Por lo general, quienes llegamos a una de estas titulaciones lo hacemos atraídos por la literatura para más tarde enamorarnos de una lengua en particular o del lenguaje en general y acabar, en muchos casos, convertidos en profesores bien de literatura, bien de lengua. Con frecuencia, estas facetas se suceden o incluso se solapan en la vida profesional de algunos de nosotros. Solo quienes se aventuran en el camino de la lingüística aplicada llegan a necesitar de la psicometría. Puesto que somos pocos los lingüistas que acabamos usándola, quienes diseñan los currículos de las titulaciones universitarias se creen obligados a priorizar disciplinas más relevantes desde otros puntos de vista. Existe otra explicación más simple para esta ausencia de la psicometría en los planes de estudio universitarios: de alguna manera, muchos de quienes nos entregamos a la literatura o la lingüística lo hacemos huyendo de las ciencias exactas, que son parte importante de la psicometría, y dado que en las facultades de lingüística las decisiones sobre el currículo las suelen tomar los «renegados» de las ciencias exactas, tiene sentido que la psicometría ocupe un lugar marginal.

Con todo, la importancia de la psicometría es capital. Para comenzar, unos conocimientos mínimos sobre psicometría dotan tanto al profesor como al lingüista de herramientas que le permiten una observación empírica y objetiva de la realidad. En la actualidad, confeccionar exámenes sin conocimientos de psicometría es como intentar hacer física sin conocimientos de matemáticas. Aunque árida, la psicometría también es necesaria. Tanto si encaminamos nuestros pasos al terreno de la lingüística aplicada como si no, la psicometría aporta a un filólogo tanta perspectiva como el estudio de la lingüística diacrónica o la literatura comparada.

La psicometría emana de la psicología, tiene una estrecha relación con la estadística, y solo se ha vinculado al estudio de las lenguas en época reciente (*vid.* Muñiz, 2010). El antecedente más cercano a la evaluación psicométrica actual lo encontramos en las universidades europeas de la Edad Media, que introdujeron la evaluación oral en el siglo XIII y la escrita en el XVI (Buchanan y Finch, 2005). Habría que esperar unos siglos hasta que el materialismo científico y las ideas darwinistas del siglo XIX pavimentaran el camino hacia una forma de psicología basada en la medición. De hecho, sería el primo de Darwin, Francis Galton, quien articularía las ideas fundamentales de la psicometría moderna (*ibid.*; Newton y Shaw, 2014:13). Los franceses Alfred Binet y Théodore Simon, coetáneos de Galton, desarrollaron el celebérrimo test de inteligencia Binet-Simon (Binet y Simon, 1948) que, posteriormente, tras ser revisado por el psicólogo Lewis Terman de la Universidad de Stanford, daría lugar al también famoso test Stanford-Binet.

Durante la Primera Guerra Mundial, los tests de inteligencia similares al Stanford-Binet se popularizaron no sin críticas. Entrada la primera mitad del siglo

XX, la psicometría consigue afianzar su base científica y se ensancha para dar cabida a diferentes corrientes (*vid*. Engelhard y Wang, 2021:1-6). En las próximas secciones hablaremos de dos de ellas, las que consideramos más relevantes para la evaluación de lenguas y que, de alguna manera, también suponen un recorrido histórico por la psicometría desde principios del siglo XX; nos referimos a la teoría clásica de los tests por un lado y a la teoría de respuesta al ítem y los modelos Rasch por otro.

2.1.1. Teoría clásica de los tests (TCT)

El adjetivo «clásica» no añade a este sistema de análisis un valor cronológico o un matiz de superioridad con respecto a la teoría de respuesta al ítem y los análisis Rasch, de los que hablaremos más adelante.

A principios del siglo XX, la TCT surge como culmen de tres razonamientos fundamentales: el reconocimiento del error en cualquier tipo de medición, la interpretación de dicho error como una variable aleatoria y, finalmente, la definición de la correlación (Traub, 1997:8). El artículo de 1904 «The proof and measurement of association between two things», de Charles Spearman (Spearman, 2010), es considerado con frecuencia como la obra fundacional de la TCT (Traub, 1997:8; Linden y Hambleton, 1997:1).

De forma muy resumida, la TCT asume que la puntuación verdadera (V) de cada candidato a nuestros exámenes (es decir, su verdadero nivel de dominio de lengua) es una combinación de la puntuación que podemos observar (X) en dicho candidato (al realizar un examen de dominio) y del error (E) intrínseco a nuestras mediciones. Este razonamiento se expresa con una ecuación de tres variables, en la que X es observable, E está distribuido aleatoriamente entre los sujetos, y V es el constructo teórico que queremos conocer:

$$V = X - E$$

La TCT, aunque muy útil y desarrollada, tiene diferentes limitaciones, como por ejemplo el hecho de que X depende del número y dificultad de los ítems, o el hecho de que la posición de un candidato con respecto al resto depende del grupo en que se encuentre (*vid*. Hambleton y Jones, 1993:38; McNamara, 1996:151–152). Esto supone que los análisis realizados mediante la TCT pueden diferir en función de los candidatos que usemos como muestra en nuestras pruebas. Dado que lo que deseamos es ser capaces de generalizar nuestros resultados, que estos dependan de la población no es un atributo positivo y es, de hecho, la principal fuente de críticas a la TCT.

Más de un siglo de investigación en TCT ha dado lugar a una pléyade de análisis de entre los cuales nos centraremos en los que consideramos más ma-

nejables y útiles para el desarrollo básico de exámenes de dominio de lengua, a saber: el error típico de medida, el coeficiente de fiabilidad, el índice de dificultad y el índice de discriminación. Los dos primeros están relacionados directamente con la fiabilidad de las puntuaciones en nuestras pruebas, mientras que los dos últimos son análisis complementarios que pueden darnos pistas sobre posibles áreas de mejora. La literatura en este campo es extensísima y el lector interesado no tendrá problemas para encontrar multitud de referencias que profundicen en estos y otros muchos análisis (*vid*. Muñiz, 1998; Green, 2013; Field, 2014). Por este motivo, en las secciones que siguen hemos obviado la mayor parte del sustrato matemático que justifica las derivaciones de ecuaciones presentadas, cuya inclusión habría sido contraria al principio de sencillez que hemos intentado mantener en este libro.

2.1.1.1. Error típico de medida (ETM) y coeficiente de fiabilidad

Como hemos visto en la sección anterior, la TCT asume que la puntuación verdadera (V) de una persona no cambia entre ocasiones, por lo que la variabilidad observada en las puntuaciones de un candidato (X) en distintas aplicaciones de la misma prueba se debería a la variabilidad aleatoria de los errores de medida (E) entre ocasiones (Prieto y Delgado, 2010:68). Este postulado queda descrito en la fórmula:

$$V = X - E$$

Si pudiésemos realizar un examen a un mismo candidato un número infinito de veces, la media de todas las puntuaciones obtenidas sería la mejor estimación de su habilidad ya que la media de los errores aleatorios sería 0. Obviamente, no podemos repetir un examen de forma indefinida, por lo que cuanto menor sea el error (E) en nuestras medidas, más cerca estará la puntuación observada (X) de la puntuación verdadera (V). Para cuantificar la influencia del error de medida en las puntuaciones de un grupo de sujetos se usan estadísticos de fiabilidad tales como la desviación típica de los errores y el coeficiente de fiabilidad.

La desviación típica de los errores, también denominada error típico de medida (ETM), indica la precisión de las puntuaciones, es decir, cuán alejadas podrían estar estas de las puntuaciones verdaderas de los candidatos (Prieto y Delgado, 2010:68). Puesto que los valores E son inobservables, no es posible calcular directamente su desviación típica. Sin embargo, el ETM puede ser estimado fácilmente a partir del coeficiente de fiabilidad (ETM S_x $(1 - R_{xx})^{1/2}$) y es facilitado por los paquetes de *software* de análisis estadístico más usados. El ETM nos sirve para estimar la precisión de las medidas y poder así trabajar con la puntuación verdadera de un candidato que, recordemos, es un valor inobservable. Dicho de otra

manera, como no podemos medir directamente la puntuación verdadera de un candidato (V), asumimos que esta estará en torno a su puntuación observada (X) en los límites que establezca el ETM, límites que podremos ensanchar o estrechar aplicando un intervalo de confianza (Z_{IC}) determinado:

$$V = X \pm (ETM \cdot Z_{IC})$$

En la TCT, la convención es que para establecer un intervalo de confianza del 90 % se multiplique por $Z_{IC} = 1.65$; para establecer un intervalo de confianza del 95 % por $Z_{IC} = 1.96$; y por $Z_{IC} = 2.58$ para intervalos del 99 %. Así, por ejemplo, ante un hipotético ETM de 0.7 y una calificación observada (X) de 6 puntos, podríamos decir que la puntuación verdadera de un candidato es 6 ± 1.155 (6 ± (0.7 x 1.65)) con un 90 % de confianza, o lo que es lo mismo, la puntuación verdadera (V) de este candidato está entre 4.845 y 7.155 con un 90 % de probabilidad. De la misma manera, si quisiésemos establecer un intervalo de confianza del 95 % (el más habitual), diríamos que la puntuación verdadera del candidato es 6 ± 1.372 (6 ± (0.7 x 1.96)), es decir, esta estaría entre 4.628 y 7.372 con un 95 % de probabilidad. Finalmente, si quisiésemos aplicar el intervalo de confianza más amplio, diríamos que la puntuación verdadera del candidato es 6 ± 1.806 (6 ± (0.7 x 2.58)) con un 99 % de confianza, o lo que es lo mismo, estaría entre 4.194 y 7.806 con un 99 % de probabilidad.

En algunos programas informáticos, el ETM viene representado por las letras SE(M), iniciales del sintagma inglés *standard error* (*of measurement*). A diferencia de lo que veremos más adelante cuando hablemos de la TRI y de Rasch, en la TCT el ETM «es una medida global del error, un único valor aplicable de forma general a todas las puntuaciones de la población» (Prieto y Delgado, 2010:69). Esto quiere decir que en la TCT se asume falsamente que el error de medida es el mismo para los candidatos que obtienen puntuaciones bajas y altas. La «valoración del ETM dependerá de la magnitud de los objetos que se midan: dos gramos es un error despreciable si se pesan objetos muy pesados como sacos de cereales, pero es un error notable si se pesan objetos más livianos como los diamantes» (*ibid.*:68), por lo que habrá que analizar este estadístico en el contexto del que se extrae. En los ejemplos descritos en el párrafo anterior, un ETM de 0.7 supone que la puntuación verdadera de un candidato con una puntuación observada de 6 pueda quedar por debajo del aprobado incluso considerando los intervalos de confianza más restrictivos.

En un momento dado de la historia de la TCT, se demostró que todas las asunciones basadas en la hipótesis de repetir un examen a un candidato en múltiples ocasiones también eran ciertas en los casos en que un mismo examen se aplica a múltiples candidatos una sola vez (Allen y Yen, 1979): los errores se

distribuyen aleatoriamente entre las personas y son independientes de las puntuaciones verdaderas. Este hallazgo acelera de forma considerable el análisis de nuestras pruebas ya que nos permite obtener estadísticos como el coeficiente de fiabilidad evaluando a un conjunto representativo de candidatos en lugar de tener que repetir la prueba infinitas veces a la misma persona.

Desde los primeros trabajos de Spearman se define el coeficiente de fiabilidad de un test como la correlación entre las puntuaciones obtenidas por un grupo de sujetos en dos medidas paralelas del mismo. Así, si un test tuviera una fiabilidad perfecta, las puntuaciones obtenidas por estos sujetos en cada una de las dos medidas paralelas deberían ser idénticas y, por tanto, la correlación entre las puntuaciones sería 1 (*vid*. Meneses *et al*., 2013:78).

De esta asunción y de los supuestos de la TCT se deriva que la varianza de las puntuaciones observadas en un grupo de candidatos (S^2_X) tiene dos componentes: la varianza de las puntuaciones verdaderas (S^2_V) y la varianza de los errores (S^2_E):

$$S^2_X = S^2_V + S^2_E$$

La varianza es la media de las diferencias al cuadrado entre cada puntuación y su media ($S^2_X = \sum(X - M_X)^2/N$), y es un concepto importante porque nos habla de la dispersión de los datos, es decir, de la variabilidad que generan nuestros ítems. La variabilidad es una característica deseable en nuestras pruebas y se define como la diferencia que existe entre las puntuaciones de las distintas personas que realizan un test. Si construyésemos un examen con ítems tan sencillos que permitiese a todos los participantes obtener la puntuación máxima, no podríamos establecer una jerarquía de candidatos ni diferenciar a quienes tienen más nivel de quienes tienen menos. Dado que asumimos que a nuestras pruebas concurrirán candidatos con distintos niveles, un buen examen de dominio habrá de ser sensible a estas diferencias de nivel que, como decimos, quedarán reflejadas en la varianza[1]. Cuanta menos varianza sea atribuible al error de medida (S^2_E), mejor para nuestras pruebas, y esto es precisamente lo que nos indica el coeficiente de fiabilidad (R_{XX}) de una prueba:

$$R_{XX} = S^2_V / S^2_X$$

Cuando se nos dice que coeficiente de fiabilidad de una prueba es de 0.80, lo que en realidad se nos está diciendo es que el 80 % de la variabilidad que genera

1. Si el objetivo es evaluar la competencia lingüística de un grupo homogéneo (por ejemplo, estudiantes que han recibido la misma instrucción), una baja varianza puede ser deseable. Esto indicaría que todos los estudiantes tienen habilidades similares y que el examen está midiendo efectivamente lo que se pretende.

nuestro examen es variabilidad verdadera, mientras que el 20 % restante es espuria. Para estimar empíricamente el coeficiente de fiabilidad se emplean diversos diseños de recogida de datos orientados a obtener las medidas paralelas a las que hemos hecho referencia unos párrafos más arriba. De entre ellos, los más conocidos son los de test-retest, formas paralelas, consistencia entre las partes de una prueba (KR-20, KR-21 y Cronbach) y consistencia de las puntuaciones de distintos calificadores (Prieto y Delgado, 2010:68). Estos coeficientes, que pueden oscilar entre 0 y 1, serán tanto mejores cuanto más se aproximen a 1. El valor de 0.7 suele considerarse como el mínimamente aceptable (*vid*. Green, 2013:38).

Programas informáticos de análisis psicométrico como los que veremos en la sección 2.2.1 operan en nuestros datos los cálculos que nos permiten centrarnos en la interpretación del ETM y el coeficiente de fiabilidad, así como en la interpretación de los índices de dificultad y discriminación, que describimos a continuación.

2.1.1.2. Índice de dificultad

El índice de dificultad (*facility value* en inglés) indica la proporción de acertantes que han respondido de forma correcta a un ítem. Para calcular esta proporción (p) se divide la frecuencia de acertantes (f_A) entre el número de casos (N):

$$p = f_A / N$$

Tomemos como ejemplo un ítem cualquiera de respuesta múltiple. Supongamos que diseñamos un ítem al que llamaremos ítem 1, para el que también diseñaremos tres respuestas. De entre estas opciones, a) es la que consideramos correcta (o «clave») mientras que b) y c) serán opciones incorrectas (o «distractores»). Imaginemos ahora que pasamos este ítem a 100 candidatos (N). De entre estos 100 candidatos, 60 eligen la opción a), 25 la opción b) y 15 la opción c). Puesto que 60 candidatos eligieron la respuesta correcta (f_A), diremos que este ítem en particular tiene un índice de dificultad (p) de 0.6 (60/100):

i. Ítem 1 respondido por los 100 candidatos del grupo A
  a) 60 candidatos (respuesta correcta)
  b) 25 candidatos
  c) 15 candidatos

El índice de dificultad, no obstante, es insuficiente por sí solo para ayudarnos a hacer inferencias relevantes sobre nuestros ítems. Por ello, debe estar acompañado del análisis de las frecuencias de los distractores. Esto, unido al índice de discriminación, nos permitirá identificar posibles errores en la clave de corrección,

distractores excesivamente fuertes o inconsistencias en el funcionamiento de los ítems.

Para entender la importancia de la frecuencia de los distractores, supongamos ahora que 150 candidatos responden a un nuevo ítem, al que llamaremos ítem 2. Observamos que 15 escogen la respuesta a), 75 escogen la b) y 60 la c), que es la correcta:

ii. Ítem 2, respondido por 150 candidatos
a) 15 candidatos
b) 75 candidatos
c) 60 candidatos (respuesta correcta)

En este caso, el índice de dificultad del ítem es 0.4 (60/150). La desigual concentración de respuestas en las distintas opciones (observemos que b) atrae incluso a más candidatos que la opción correcta) podría estar apuntando a una codificación incorrecta de la clave (es decir, la respuesta correcta es b) y no c)), podría estar indicándonos que el distractor b) es demasiado atractivo (quizás es muy similar la respuesta correcta o quizás también es una respuesta correcta) o podría indicar, sencillamente, que el ítem 2 es más difícil que el ítem 1. El análisis de distractores es, como vemos, de vital importancia y, con frecuencia, una importante fuente de información sobre nuestros ítems que puede obtenerse de forma sencilla (Green, 2013:185-193).

2.1.1.3. Índices de discriminación: Kelley y correlación ítem-test

Otro estadístico muy útil dentro de la TCT es el índice de discriminación u homogeneidad. Un buen ítem ha de ser como un bisturí, es decir, ha de separar inequívocamente a aquellos candidatos que poseen el nivel del ítem de los que no. El índice de discriminación es el guarismo que nos indica cuán afilado es nuestro bisturí, es decir, cuán bueno es discriminando entre los candidatos que tienen el nivel y los que no.

Técnicamente hablando, el índice de discriminación refleja la medida en que el éxito en un ítem se corresponde con el éxito global en la prueba (Kelley *et al.*, 2002:883). Existen distintas formas de calcular este estadístico. En esta sección comenzaremos describiendo la fórmula propuesta por Kelley (1939), que nos ayudará a conceptualizar la idea de discriminación, y más tarde definiremos los estadísticos basados en la correlación ítem-test que aportan los programas de análisis psicométrico.

Kelley (1939) estableció que el índice de discriminación se podía obtener estudiando al 27 % de candidatos de habilidad superior y al 27 % de candidatos de habilidad inferior. Engelhart (1965) demostró posteriormente que la fórmula de

Kelley (1939) también funcionaba comparando el tercio de candidatos superior con el tercio inferior. Lo que Engelhart (1965) propone es dividir en tres grupos de igual tamaño a los candidatos que responden a un ítem particular. En el subgrupo 1 (SG1) se ubica al tercio de candidatos con mejores puntuaciones globales en la prueba, en el subgrupo 3 (SG3) se ubica al tercio de candidatos con peores calificaciones y en el subgrupo 2 (SG2) al resto. Tras dividir a los candidatos, se resta al número de candidatos del SG1 que respondieron al ítem 1 correctamente el número de candidatos del SG3 que respondieron correctamente al mismo ítem. Los candidatos del SG2 no se usan en el cálculo. Finalmente, dividimos el resultado de la resta entre el número de candidatos de los grupos (que, como hemos dicho, será un tercio del total). El resultado, un valor entre -1 y +1, nos dirá cuán afilado es nuestro bisturí. Veamos un ejemplo.

Imaginemos que 300 candidatos responden a nuestro ítem 1. Puesto que los tres subgrupos en que dividamos a estos 300 candidatos han de ser iguales, cada uno incluirá a 100 personas. Como antes hemos indicado, en el SG1 ubicaremos a los 100 candidatos con mejor calificación global y en el SG3 a los 100 candidatos con peor calificación global. Obviamos a los candidatos del SG2. El siguiente paso es investigar cuántos de los candidatos del SG1 han respondido correctamente al ítem 1. Dado que son los de mejores calificaciones, cabe esperar que muchos de ellos lo respondan correctamente. Supongamos que todos los candidatos del SG1 respondieron correctamente al ítem 1 mientras que solo 30 candidatos del SG3 lo respondieron correctamente (*vid*. Martínez, 2011:68). Tras restar al número de candidatos del SG1 el número de candidatos del SG3 dividimos el resultado por un tercio del número total de candidatos:

Ítem 1 (n = 300 candidatos)	**Respuestas correctas**
SG1 (los 100 candidatos con mejor calificación global)	**100** candidatos respondieron el Ítem 1 correctamente
SG2 (los 100 candidatos con calificación intermedia)	No se utiliza en el cálculo
SG3 (los 100 candidatos con peor calificación global)	**30** candidatos respondieron el Ítem 1 correctamente
$\frac{SG1-SG3}{\frac{n}{3}} = \frac{100-30}{\frac{300}{3}} = \frac{70}{100} = 0.7$	

Tabla 1. Cálculo del índice de discriminación de un ítem

Modificando ligeramente los datos de la tabla 1, podemos ejemplificar una de las limitaciones de la TCT mencionada en la sección 2.1.1. Si utilizamos el mismo ítem 1 con un nuevo grupo de 300 candidatos de habilidad más uniforme y obtenemos valores de, digamos, SG1 = 89 y SG3 = 50, el índice de discriminación baja hasta .39. Así, el mismo ítem podría pasar de excelente a ser considerado como razonablemente bueno simplemente porque se probó en un grupo de candidatos de un nivel distinto.

Para compensar estas limitaciones, la TCT ha generado formas más robustas de calcular el índice de discriminación, por ejemplo, la correlación ítem-test. La correlación ítem-test indica la magnitud (entre 0 y 1) y el sentido (positivo o negativo) de la asociación entre las puntuaciones de los candidatos en un ítem (i) y la puntuación total en la prueba (X). Si las puntuaciones del ítem son dicotómicas (1/0, acierto/error) se suele utilizar la correlación biserial puntual para cuantificar esta asociación. Así, si los candidatos que aciertan el ítem obtienen mayores puntuaciones X que los que fallan, la correlación sería alta y positiva. Si los que aciertan obtienen peores puntuaciones que los que fallan, la correlación sería alta y negativa. En la práctica, los ítems con una discriminación aceptable han de tener índices de discriminación mayores de +0.25 o +0.30. Estos índices indicarían que el ítem funciona de manera homogénea con las puntuaciones en la prueba, un indicador de que el ítem mide los mismos atributos que el resto de la prueba (Prieto, comunicación personal).

El primero de estos índices de discriminación ítem-test es r_{iX}, un estadístico «inflado» por la autocorrelación puesto que correlaciona la parte (el ítem) con el todo al que pertenece (la puntuación total). En la columna *PTMEASUR-AL CORR.* de la tabla 2, extraída de *Winsteps* (Linacre, 2024b), se muestran los índices de discriminación sin corregir de un conjunto de ítems.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	**PTMEASUR-AL CORR.**	EXACT MATCH EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
1	126	131	-2.79	.49	1.16	.5	.98	.2	.20	.27	95.4	96.3	CL_1_ISE
2	128	131	-3.39	.61	.83	-.2	.23	-1.0	.35	.22	97.7	97.7	CL_2_ISE
3	120	131	-1.81	.34	1.11	.5	.93	.0	.28	.34	90.8	92.2	CL_3_CCR
4	101	131	-.38	.23	.82	-1.5	.63	-1.7	.57	.43	82.4	80.2	CL_4_INI
5	85	131	.38	.21	.86	-1.6	.74	-1.7	.56	.45	76.3	73.4	CL_5_ISP
6	54	131	1.61	.20	1.20	2.4	2.37	6.5	.22	.44	67.9	69.9	CL_6_IIP
7	118	131	-1.59	.32	.87	-.5	.72	-.5	.45	.36	90.8	90.9	CL_7_ISE
8	90	131	.16	.21	.93	-.7	.94	-.3	.49	.45	77.9	75.2	CL_8_CCR
9	113	131	-1.15	.28	1.00	.1	.74	-.7	.42	.39	87.0	87.6	CL_9_ISP
10	97	131	-.17	.22	.73	-2.5	.56	-2.4	.64	.43	85.5	78.2	CL_10_IN
11	79	131	.63	.20	1.12	1.4	1.29	1.9	.34	.45	69.5	71.5	CL_11_IN
12	102	131	-.43	.24	.98	-.1	.81	-.8	.46	.42	78.6	80.7	CL_12_IS
13	87	131	.29	.21	1.06	.7	.92	-.4	.43	.45	70.2	74.0	CL_13_II
14	74	131	.83	.20	1.15	1.8	1.11	.9	.35	.45	64.1	70.2	CL_14_II
15	83	131	.46	.20	1.00	.1	.91	-.5	.46	.45	71.0	72.7	CL_15_IS

Tabla 2. Índices de discriminación sin corregir de un conjunto de ítems

Para evitar esta autocorrelación, especialmente cuando la prueba tiene pocos ítems, se suele usar la correlación corregida $r_{i(X-i)}$, en la que se resta en cada sujeto la puntuación del ítem. Gracias a esta corrección, los valores de r_{ix} siempre son mayores que los de $r_{i(X-i)}$, si bien la diferencia es pequeña cuando el número de ítems es mayor de veinte. Muchos programas permiten elegir entre uno u otro índice de discriminación o aportan ambos (Prieto, comunicación personal). En el caso de *SPSS* (IBM Corporation, 2023), por ejemplo, cuando realizamos un análisis de fiabilidad para obtener el alfa de Cronbach, la Correlación total de elementos corregida nos muestra $r_{i(X-i)}$, como se observa en la tabla 3.

	Media de escala si el elemento se ha suprimido	Varianza de escala si el elemento se ha suprimido	Correlación total de elementos corregida	Alfa de Cronbach si el elemento se ha suprimido
CL_1_ISEx_B	19,37	30,390	,173	,838
CL_2_ISEx_B	19,36	30,232	,327	,837
CL_3_CCRR_C	19,42	30,030	,225	,837
CL_4_InIS_C	19,56	28,217	,535	,829
CL_5_ISPC_A	19,69	27,894	,528	,828
CL_6_IIP_A	19,92	29,810	,137	,842
CL_7_ISEC_B	19,44	29,294	,434	,833
CL_8_CCRR_C	19,65	28,245	,471	,830
CL_9_ISPC_B	19,47	29,144	,410	,833
CL_10_InIS_D	20,34	30,794	,000	,840
CL_11_InIS_A	19,73	29,074	,279	,837
CL_12_ISP_B	19,56	28,772	,413	,832
CL_13_IIPt_C	19,67	28,776	,353	,834
CL_14_IIP_C	19,77	29,147	,260	,838
CL_15_ISPC_C	19,70	28,580	,383	,833

Tabla 3. Índices de discriminación corregidos de un conjunto de ítems

Las tablas 2 y 3, pues, muestran los valores r_{iX} y $r_{i(X-i)}$ del mismo conjunto de ítems, concretamente un conjunto de ítems diseñado en la Unidad de Evaluación y Certificación en Lenguas de la Escuela de Idiomas de la Universidad de Antioquia, Colombia. Como se observa, los valores de r_{iX} son mayores y, aunque existen distintos criterios a la hora de interpretar estos estadísticos (Green, 2013:29), está generalmente aceptado que:

Índice de discriminación	Interpretación
.40 o más	Ítem muy bueno
entre .30 y .39	Razonablemente bueno, pero con margen de mejora
entre .20 y .29	Ítems marginales que necesitan mejora
.19 o menos	Malos ítems que deben ser rechazados o revisados

Tabla 4. Interpretación del índice de discriminación de un ítem

Cabe aquí hacer una referencia a aquellos ítems con índices de discriminación negativos. En el caso de la fórmula de Kelly (1939), aparecerán índices de discriminación negativos siempre que el SG1 tenga un valor inferior al del SG3, es decir, siempre que haya menos candidatos que respondan bien el ítem en el grupo de los que mejor calificación global obtuvieron. Un resultado de este tipo no tiene sentido, dado que lo esperable siempre es que el ítem lo respondan bien más candidatos dentro del grupo con más habilidad global. Si, por ejemplo, usamos los valores SG1 = 40 y SG3 = 55, el índice de discriminación bajaría a -.15. En el caso de r_{iX} y $r_{i(X-i)}$ aparecerán correlaciones negativas cuando la media en el test de los que fallan el ítem es mayor que la media de los que lo aciertan. Generalmente esto ocurre cuando el ítem es confuso (puede que haya dos respuestas muy parecidas y que los mejores candidatos tiendan a una de ellas y el resto a otra) o, sencillamente, cuando la clave no es la correcta, es decir, cuando hemos tomado como correcta una respuesta incorrecta.

Algunas fuentes (Abad *et al*. 2011) denominan índice de fiabilidad del ítem al producto de la desviación típica del ítem por su índice de discriminación ($s_i r_{ix}$). Este estadístico indica la contribución del ítem al coeficiente alfa de Cronbach (*vid*. 2.1.1.1). No obstante, los índices de dificultad y discriminación de los que hemos hablado hasta este punto bastan para analizar y seleccionar los ítems adecuados para nuestra prueba.

2.1.2. Teoría de respuesta al ítem (TRI) y modelos Rasch

En la sección 2.1.1 hemos explicado las características de la TCT y propuesto algunos ejemplos de análisis. En la sección 2.1.2 haremos lo propio con los modelos

pertenecientes a la TRI (también conocida como IRT por su nombre en inglés *item response theory*, como teoría moderna de tests, y como teoría de variable latente), y con los modelos Rasch. Unos y otros tienen su origen en los trabajos de Thurstone (1928), quien comenzó a pavimentar un camino distinto al propuesto por la TCT con el objeto de superar las limitaciones de esta última. Birnbaum ensanchó este camino, y autores como Rasch (1980) y Lord (1980) lo continuaron transitando ramales con características propias. Para una reseña histórica más detallada, véase Muñiz (2010).

Dado que todos estos modelos matemáticos parten de un tronco común, la diferencia entre unos y otros no siempre es clara. Algunos autores, por ejemplo, incluyen dentro de la TRI a los modelos Rasch mientras que otros aluden a cuestiones filosóficas para considerar a estos últimos como una categoría independiente (Engelhard y Wang, 2021:8). Ciertamente, todos ellos comparten la característica de ser modelos probabilísticos. Así, mientras que en la TCT la noción central, como hemos visto, es la puntuación verdadera en un examen, la noción central de la TRI y los modelos Rasch es la modelización métrica de la respuesta de una persona a un ítem dicotómico (por ejemplo, la respuesta correcta a un ítem de respuesta múltiple o la respuesta «sí» en un cuestionario) (*vid*. Hambleton *et al*., 1991). En román paladino, esto quiere decir que mediante los modelos TRI y Rasch podemos calcular la probabilidad que un candidato tiene de acertar un ítem. Veamos qué significa esto con un ejemplo tomado de Verhelst (2004:1–2).

Imaginemos que diseñamos tres ítems con niveles ascendentes de dificultad. Llamaremos a estos ítems *i*, *j* y *k*, y los ubicaremos en una línea imaginara que representa nuestra variable latente (el dominio de un idioma). Situaremos al ítem *i* (el más fácil) a la izquierda, al ítem *j* (de dificultad intermedia) en el centro, y al ítem *k* (el más difícil) a la derecha, dividiendo así la línea de la variable latente en cuatro segmentos. Tras administrar estos ítems a distintos candidatos, observamos patrones de respuesta como los reflejados en el escalograma de la tabla 5, donde 0 indica «respuesta incorrecta» y 1 indica «respuesta correcta»:

Candidato	Ítem *i*	Ítem *j*	Ítem *k*
Candidato 1	0	0	0
Candidato 2	1	0	0
Candidato 3	1	1	0
Candidato 4	1	1	1

Tabla 5. Escalograma

Dado el patrón de respuestas obtenido, en la línea imaginaria de nuestra variable latente ubicaremos al candidato 1 a la izquierda del ítem *i* (el más fácil) puesto que no ha respondido correctamente a ninguno de los ítems. Es decir, teorizamos que el nivel del candidato 1 es inferior al nivel del ítem más sencillo porque no lo ha respondido correctamente. En cambio, al candidato 4 lo ubicaremos a la derecha del ítem *k* (el más difícil) puesto que ha respondido bien a todos los ítems. En este caso, asumimos que el nivel del candidato 4 es superior al del ítem más difícil. Al candidato 2 lo ubicamos en el segmento acotado por los ítems *i* y *j* (acertó el *i* y falló el *j*), y al candidato 3 entre los ítems *j* y *k* (acertó el *j* y todo lo que estaba a la izquierda de este, pero falló el *k*). Partiendo de esta base, si nos encontrásemos con un hipotético candidato 5 de habilidad similar a la del candidato 3 podríamos anticipar que la probabilidad de que responda al ítem *k* es baja pero que la probabilidad de que responda al ítem *i* es alta. Un candidato de habilidad similar a la del candidato 1 tendría una probabilidad muy baja de responder correctamente al ítem *k*, etc.

De esta forma, se establece una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de acertar estos (Muñiz, 2010:64), es decir, se puede crear una función matemática de valor predictivo. Para entender lo que son las funciones matemáticas, es útil conceptualizarlas como máquinas que tienen un puerto de entrada, otro de salida, y a las que se le pueden asignar tareas de computación. Así, si asignamos a una función la tarea de «multiplicar por dos» y le damos el número 3 como entrada, obtendremos el valor 6 como salida. Si le asignamos la tarea «elevar al cuadrado» y le suministramos el valor 4 como entrada, obtendremos el 16 como salida, etc. La función característica de la TRI nos permite computar hasta tres parámetros además de la habilidad de los candidatos: la dificultad de los ítems, el índice de discriminación de estos y la probabilidad de acertar el ítem al azar:

$$P_{ij}(X = 1) = c_i + (1 - c_i)\frac{\exp[a_i(\theta_i - b_i)]}{1 + \exp[a_i(\theta_i - b_i)]}$$

donde

P_{ij} = probabilidad de que el sujeto *j* acierte el ítem *i*
b_i = dificultad del ítem *i*
a_i = discriminación del ítem *i*
c_i = probabilidad de acertar el ítem *i* al azar
θ_i = puntuación en la variable medida
exp = función exponencial

Cuando en la función se incluyen los tres parámetros mencionados (dificultad del ítem, discriminación y azar), se habla del modelo logístico de tres parámetros (3PL). Cuando la probabilidad de azar no se tiene en consideración ($c_i = 0$), se habla del modelo de dos parámetros (2PL) y, finalmente, cuando la discriminación se considera constante ($a_i = 1$), estamos ante el modelo de un parámetro (1PL). Es precisamente este último el que algunos expertos equiparan al modelo Rasch, que se expresa independientemente de la siguiente manera:

$$\Phi_{ni1} = \frac{\exp(\theta_n - \delta_{i1})}{1 + \exp(\theta_n - \delta_{i1})}$$

donde

Φ_{ni1} = probabilidad de respuesta correcta del candidato n en el ítem i
θ_n = habilidad del candidato n
δ_{i1} = dificultad del ítem i
exp = función exponencial

Mediante estas funciones (recordemos: máquinas de computación que admiten valores de entrada y generan otros de salida) obtenemos uno de los elementos gráficos más conocidos de la TRI y los modelos Rasch, las Curvas Características del Ítem (CCI), que ilustramos en la figura 1. Las curvas de la figura 1 nos indican que un candidato de habilidad -2 (valor de entrada) tendrá una probabilidad del 60 % (valor de salida) de responder correctamente al ítem *i*, o que para tener un 60 % de probabilidad de acierto en los ítems *j* y *k* se necesitan habilidades de 0 y +2 respectivamente. Así, puesto que para tener la misma probabilidad de éxito en los distintos ítems se necesitan habilidades cada vez mayores, se confirma el orden de dificultad teórico con el que habíamos diseñado nuestros ítems (el *i* es el más sencillo, el *j* es el de dificultad intermedia y el *k* es el más difícil). De la misma manera, conocida la habilidad de cualquier otro candidato (valor de entrada) podríamos ubicarlo en el eje horizontal y trazar una línea vertical para ver en qué punto corta las curvas y establecer así la probabilidad de acierto que este tiene en uno u otro ítem (valor de salida).

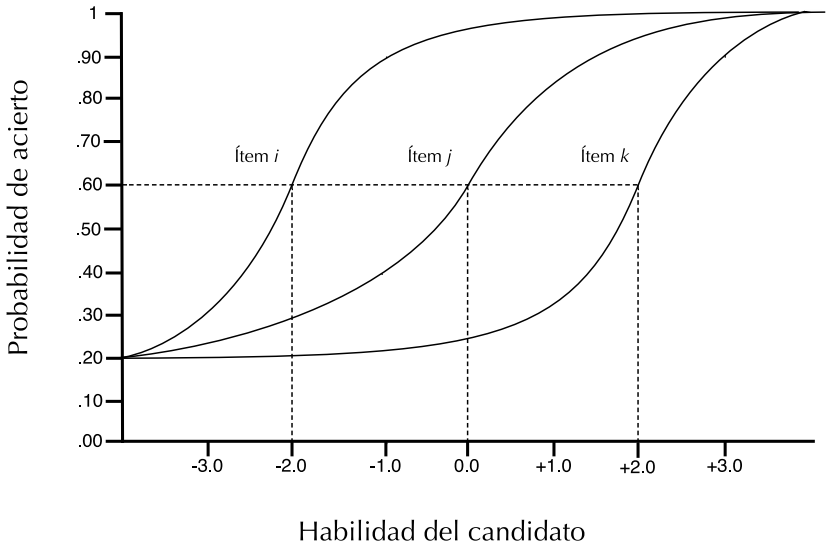


Figura 1. Curvas de probabilidad

Además de su naturaleza probabilística, la familia de modelos Rasch y el modelo TRI 1PL poseen una característica diferenciadora con respecto a otros modelos TRI, la llamada «invarianza» u «objetividad específica» (Eckes, 2009:4). Esta característica permite la comparación entre individuos sin que los instrumentos utilizados (ítems, tareas, exámenes, escalas, etc.) influyan en la medición y, de la misma manera, permite comparar ítems sin interferencias provenientes de las características de los candidatos (Rasch, 1980:xx; Stenner, 1994:374). La invarianza u objetividad específica nos permitirá, por ejemplo, ubicar a individuos de diferentes grupos en una misma escala aun cuando estos hayan respondido a ítems distintos (Reise *et al.*, 2005:100). La objetividad específica también es interesante a la hora de eliminar las limitaciones que encontrábamos en la TCT, por ejemplo, al calcular los índices de discriminación que, como veíamos en la sección 2.1.1.3, podían variar según la muestra. Si se piensa bien, esta invarianza u objetividad específica es una cualidad fundamental para cualquier examen de dominio toda vez que una buena herramienta de medida (ítems, tareas, exámenes, escalas, etc.) no debe ver comprometida su función por el elemento que intenta medir, ya que esto puede afectar o incluso invalidar las medidas obtenidas (*vid.* Thurstone, 1928:547) y, por ende, las inferencias de ellas extraídas. De acuerdo con Engelhard y Wang

(2001:6), y con Engelhard y Wind (2018:10), en los modelos de medición invariante como el modelo Rasch:

Medidas de las personas

1. La medición de las personas debe ser independiente de cualesquiera ítems que se usen para la medición.
2. Una persona más capaz debe tener siempre una probabilidad más alta de éxito en un ítem que una persona menos capaz.

Calibración de los ítems

3. La calibración de los ítems debe ser independiente de las personas que se usen para su calibración.
4. Cualquier persona debe tener una probabilidad más alta de éxito en un ítem fácil que en un ítem difícil.

Mapa de Wright (mapa de la variable)

5. Tanto las personas como los ítems deben poder ser ubicados en un único continuo latente.

Como hemos comentado anteriormente, el modelo Rasch original es dicotómico, es decir, tan solo es sensible a respuestas del tipo «correcto» frente a «incorrecto» (o a respuestas mutuamente excluyentes, por ejemplo, dentro de un ítem de respuesta múltiple), y no es aplicable a respuestas «casi correctas» o «parcialmente correctas» (*vid*. Wright y Masters, 1982:38). Esta forma básica del modelo Rasch sería posteriormente expandida a fórmulas más sofisticadas para respuestas politómicas, es decir, aquellas que admiten tres o más respuestas (como por ejemplo las escalas Likert o las escalas evaluativas que utilizamos para la corrección de la producción oral o escrita de nuestros exámenes). De entre todos los modelos politómicos de la familia Rasch los más conocidos son el Modelo de Escalas de Calificación (Andrich, 1978), el Modelo de Crédito Parcial (Masters, 1982) y el Many-Facet Rasch Measurement (MFRM) (Linacre, 1989). El modelo politómico MFRM, que puede considerarse como una extensión del modelo de Crédito Parcial o incluso del modelo dicotómico siempre que se incluyan más de dos facetas en el marco de medición, «es aplicable a los casos en los que existen diversas facetas de la medición (personas, ítems, calificaciones, categorías ordenadas de evaluación, etc.) que pueden contribuir al error de medida» (Prieto, 2011:234). El modelo MFRM, pues, será particularmente útil cuando tengamos que analizar el

desempeño de nuestros candidatos mediante escalas, como veremos en el capítulo 4, y se formula de la siguiente manera:

$$\log\left(\frac{P_{nijk}}{P_{nij\,(k\text{-}1)}}\right) = B_n - D_i - C_j - F_{jk}$$

donde

P_{nijk} = la probabilidad de que una persona n reciba la calificación k en el ítem i por el calificador j
$P_{nij\,(k\text{-}1)}$ = la probabilidad de que una persona n reciba la calificación inferior (k-1) en el ítem i por el calificador j
B_n = nivel en la variable latente de la persona n
D_i = dificultad del ítem
C_j = severidad del calificador j, y
F_{jk} = localización en la variable del *paso* entre las categorías adyacentes k y k - 1 en el calificador j.

Como Prieto (2011:234) indica, en esta función

> El *logit* (log ($P_{nijk}/P_{nijk\,(k\text{-}1)}$)) es la variable dependiente y las diversas facetas (personas, ítems, calificadores, etc.) son las variables independientes. Es decir, el modelo especifica que la probabilidad de que el calificador j otorgue a una persona n una calificación (k) en lugar de la inferior (k - 1) en el ítem i depende de los efectos aditivos de la dificultad del atributo (D_i), de la severidad del calificador (C_j), del nivel de ejecución de la persona (B_n) y del valor del paso entre las categorías k y 1 (F_{jk}). En esta formulación del MFRM se asume que los parámetros de cada faceta pueden ser estimados independientemente del resto de facetas en una escala común […]. La escala *logit* puede oscilar entre 0 ± ∞. El punto 0 se fija convencionalmente en el nivel medio de los ítems, los calificadores y las categorías, permitiendo la variación libre en la escala común de las personas evaluadas.

Veremos ejemplos de análisis realizados mediante MFRM en el capítulo 4. El lector interesado en el sustrato matemático de este último párrafo y del resto de formulaciones de esta sección puede consultar Thurstone (1928), Andrich (1978), Wright y Stone (1979), Lord (1980), Rasch (1980), Wright y Masters (1982), Masters (1982), Linacre (1989), McNamara (1996), Eckes (2009), Prieto (2011), Engelhard y Wang (2021) o Engelhard y Wind (2018) entre otros.

2.1.2.1. *Logits*

Un concepto central que hemos de conocer sobre la TRI y los modelos Rasch es el de *logit*, que ya ha sido mencionado en la sección anterior. Este término procede de la contracción del sintagma inglés *logg-odds unit* que, de una forma un tanto libre, podríamos traducir como «unidades logarítmicas de probabilidad». De igual manera que la electricidad se mide en amperios, la resistencia se mide en ohmios o la distancia entre las estrellas en años luz, los *logits* son la unidad de medida en los análisis estadísticos realizados mediante TRI y Rasch.

Los valores de los *logits* pueden ir desde menos infinito hasta más infinito, si bien suelen estar entre -4 y +4. Un valor positivo de *logit* indica que el ítem es más difícil que el promedio de ítems de la prueba, mientras que un valor negativo indica que es más fácil. En el caso de los candidatos, los valores positivos identifican a quienes son más hábiles que la media, mientras que los negativos señalan a los menos hábiles que la media. Los *logits* poseen, además, otras características interesantes: son aditivos (i.e., los intervalos entre dos *logits* son siempre iguales); nos permiten ubicar en una misma línea recta o continuo latente a los diferentes ítems y candidatos observados (Wright, 1993:288) y denotan probabilidad. Es importante saber que una escala de *logits* es calculada individualmente para cada análisis realizado, por lo que, si comparamos los *logits* del examen A con los del examen B, estaremos observando dos historias distintas que no son directamente comparables. Para establecer este tipo de comparaciones, características de la medición invariable, son necesarios procedimientos adicionales de anclaje vertical u horizontal (Wolfe, 2004).

Las escalas de *logits* las calcula cualquier programa informático utilizado para este tipo de análisis y son visibles en la práctica totalidad de tablas que dichos programas generan, por lo que es importante familiarizarse con el término y su significado. Si no se comprenden de forma clara las unidades utilizadas y el significado de las medidas de nuestro análisis es virtualmente imposible implementar cualquier plan de acción basado en el marco teórico del que estas se derivan (*vid.* Engelhard y Wang, 2021:14). Para una conceptualización más técnica y precisa de los *logits* y su cálculo véanse Wright (1993:288), Hambleton *et al.* (1991), McNamara (1996:165) y especialmente Bond y Fox (2007:24–26).

2.1.2.2. Valores de ajuste al modelo

Para alcanzar la deseada medición invariante u objetividad específica es necesario que se cumplan determinados requisitos. Si estos requisitos no se dan, no podremos asegurar que las inferencias hechas a partir de nuestros análisis sean correctas y, por tanto, la validez de la prueba estará en entredicho. Así pues, tras recopilar los datos necesarios sobre nuestros ítems y candidatos hemos de pasarlos por el

tamiz de los índices de ajuste para comprobar si las respuestas observadas difieren o no de las respuestas esperadas por los modelos matemáticos.

Los valores de ajuste pueden verse como una especie de «control de calidad» sobre la información que los datos generados por nuestros exámenes aportan al análisis. Este control de calidad se realiza analizando los valores *infit* y *outfit mean square* (en adelante, valores de ajuste al modelo), que son exclusivos de los modelos Rasch. Para una descripción detallada de su cálculo e interpretación puede consultarse Engelhard y Wang (2021:37-42). Para conceptualizarlos de forma intuitiva se los puede considerar como índices que indican si los datos con los que hemos alimentado nuestro análisis (i.e. las respuestas de los candidatos a los diferentes ítems, la forma en que hemos usado una escala para puntuar a los examinandos, etc.) son útiles para la medición mediante el modelo Rasch o si, por el contrario, están distorsionando nuestras interpretaciones (por ejemplo, si las respuestas de nuestros candidatos fueron demasiado aleatorias o demasiado predecibles). Cualquier programa informático que permita el análisis Rasch calcula los índices de ajuste al modelo y, por tanto, nuestra labor principal consiste en saber si dichos valores están dentro de los rangos aceptables.

Linacre (2024b) compara los valores de ajuste al modelo con la música. Nuestros exámenes vendrían a ser una especie de partitura que ha de ser interpretada al unísono por candidatos e ítems. Que los índices de ajuste de algún ítem o candidato estén por encima o por debajo de lo recomendado significa que estos están generando un ruido que nos impide escuchar la música de forma correcta. Solo cuando los datos estén dentro de los rangos adecuados escucharemos la música con claridad. El valor deseable en los índices de ajuste al modelo es 1 (Engelhard y Wang, 2021:40) y, de acuerdo con Linacre (2024b), hemos de interpretar los índices distintos a 1 de la siguiente manera:

>2.0 – Datos que distorsionan o degradan la herramienta de medida
1.5–2.0 – Datos improductivos para la medición pero que no la degradan
0.5–1.5 – Datos productivos para la medición ¡Música celestial!
<0.5 – Datos menos productivos para la medición, que no la degradan pero que pueden producir fiabilidades y separaciones engañosamente buenas

Aunque los rangos deseables para los índices de ajuste al modelo son iguales para *infit* y *outfit mean square*, ambos tienen sensibilidades distintas y nos pueden contar historias diferentes. En general, los valores *outfit* son más sensibles a los patrones de respuestas improbables (por ejemplo 1100000001 en un examen de dificultad ascendente), mientras que los valores *infit*, más robustos, indican patrones de respuesta inconsistentes (por ejemplo 1101100101 también en un examen de dificultad ascendente). Según Engelhard y Wang (2021:39), los valores *infit* son más sensibles a patrones irregulares cuando personas e ítems están bien ubicadas en el

continuo latente, mientras que los valores *outfit* son mejores para la detección de datos improbables dentro del conjunto de la prueba. Además, (*ibid.*:31) los valores *infit* son más sensibles a respuestas inesperadas en ítems de dificultad similar a la de la habilidad del candidato, mientras que los valores *outfit* lo son a respuestas inesperadas en ítems de dificultad distinta a la de la habilidad del candidato.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	**MEASURE**	MODEL S.E.	**INFIT** MNSQ	INFIT ZSTD	**OUTFIT** MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
18	9	82	2.25	.41	1.04	.2	1.40	.7	.43	.43	88.9	90.3	CAC7_C
16	12	82	1.81	.36	.97	-.1	.81	-.2	.48	.46	87.7	87.3	CAC5_C
14	15	82	1.45	.33	1.01	.1	.80	-.3	.48	.48	85.2	84.8	CAC3_C
20	18	82	1.14	.31	.83	-1.0	.56	-1.1	.60	.49	84.0	82.4	CAC9_C
15	20	82	.95	.30	.87	-.8	.77	-.5	.57	.50	86.4	81.1	CAC4_C
19	23	82	.69	.29	.92	-.5	.69	-1.0	.57	.51	76.5	79.2	CAC8_C
3	24	82	.60	.29	.80	-1.4	1.01	.1	.59	.51	87.7	78.6	CAA3_B
2	25	82	.52	.28	.93	-.4	.95	-.1	.54	.51	82.7	78.0	CAA2_B
13	28	82	**.29**	.28	1.03	.3	.93	-.2	.51	.52	74.1	76.1	**CAC2_C**
11	29	82	**.21**	.27	1.03	.3	.92	-.3	.52	.52	69.1	75.5	**CAB6_B**
10	30	82	.14	.27	1.19	1.5	1.31	1.3	.41	.52	72.8	75.1	CAB5_C
6	34	82	-.15	.27	1.45	3.4	**1.64**	2.6	.26	.52	61.7	73.7	**CAB1_C**
5	38	82	-.43	.26	1.14	1.2	1.18	.9	.44	.52	66.7	73.0	CAA5_A
9	40	82	-.57	.26	1.18	1.5	1.48	2.2	.40	.52	64.2	72.7	CAB4_B
8	43	82	-.77	.26	.99	-.1	.92	-.3	.53	.52	75.3	72.6	CAB3_C
7	49	82	-1.18	.27	.99	.0	.93	-.3	.51	.50	74.1	73.8	CAB2_A
4	54	82	-1.54	.27	.86	-1.1	1.05	.3	.54	.49	84.0	75.3	CAA4_C
17	54	82	-1.54	.27	.72	-2.4	.56	-1.9	.65	.49	84.0	75.3	CAC6_C
1	58	82	-1.85	.28	.89	-.8	1.22	.8	.50	.47	80.2	77.0	CAA1_B
12	60	82	-2.01	.29	.94	-.4	1.04	.2	.49	.46	80.2	78.3	CAC1_C
MEAN	33.2	82.0	.00	.29	.99	.0	1.01	.1			78.3	78.0	
P.SD	15.3	.0	1.20	.04	.16	1.2	.28	1.0			8.0	4.8	

Tabla 6. Valores de ajuste al modelo de los ítems

En la tabla 6 mostramos un ejemplo de índices de ajuste correspondientes al análisis de una prueba de comprensión lectora realizada en el Centro de Estudios Avanzados en Lenguas Modernas de la Universidad de Jaén, España. Los índices de ajuste de estos ítems aparecen en las columnas *INFIT MNSQ* y *OUTFIT MNSQ* respectivamente. Como indicábamos anteriormente, el valor óptimo para estos estadísticos es 1 y, si su valor es distinto a 1, este ha de estar en el rango que va desde .5 a 1.5 *logits*. De entre todos nuestros ítems, tan solo el ítem CAB1_C tiene un *outfit mean square* mayor de 1.5, concretamente de 1.64, y, por lo tanto, es el único que deberíamos investigar, ya que parece estar mostrando una desviación moderada entre lo observado y lo predicho por el modelo. El resto de ítems están dentro de los valores de referencia y, por tanto, la información que suministran es productiva para la interpretación de los resultados. Vemos que todos nuestros ítems están ejecutando una partitura que se escucha con claridad y sin distorsiones. En caso de que fuese necesario, podríamos llevar a cabo este mismo tipo de análisis con los candidatos si, por ejemplo, estuviésemos más interesados en analizar el desempeño de estos que en analizar la calidad técnica de nuestros ítems.

La tabla 6 también servirá de complemento al análisis del mapa de la variable, del que hablaremos en la próxima sección. Esta tabla ordena los ítems según su dificultad, que aparece en la columna *MEASURE*. La dificultad de los ítems va desde la del más difícil, el CAC7_C, con 2.25 *logits* hasta el más sencillo, el CAC1_C con -2.01 *logits*. Si observamos la figura 2 de más abajo, veremos que, efectivamente, CAC7_C está en la parte más alta de la mitad derecha y CAC1_C en la parte más baja. Gracias a la tabla 6 podemos analizar con más detalle pequeñas diferencias entre algunos de los ítems. Por ejemplo, en la figura 2 los ítems CAB6_B y CAC2_C aparecen en paralelo, como si tuvieran el mismo nivel de dificultad. Sin embargo, gracias a la tabla 6 apreciamos que el primero tiene una dificultad de .21 mientras que la del segundo es de .29, por lo que CAC2_C es ligeramente más difícil.

2.1.2.3. Mapa de la variable

El mapa de la variable es una herramienta gráfica muy útil que nos permite ubicar tanto a los candidatos como a los ítems de nuestros exámenes de dominio a lo largo del mismo continuo latente. Como vimos en la sección 2.1.2, la posibilidad de ubicar a unos y otros en una misma línea es una característica de la medición invariante u objetividad específica y, por tanto, este tipo de representaciones tan solo son posibles en los análisis realizados mediante los modelos 1PL o Rasch. El mapa de variables obtiene su nombre de la voz inglesa *variable map* y también es conocido como *Wright map* en honor al psicometrista norteamericano Ben Wright, que hizo numerosas contribuciones a este sistema de análisis. Veamos un ejemplo y cómo se interpreta.

A la izquierda de la figura 2 encontramos la escala de *logits*, que va desde -4 hasta +3. El *logit* 0 de esta escala se sitúa donde está la habilidad media de los ítems. El eje de líneas verticales del centro separa a los candidatos, que quedan a la izquierda, de los ítems a los que estos respondieron, que quedan a la derecha. Cuanto más arriba se ubiquen, más hábiles los candidatos y más difíciles los ítems. Como vemos, cada candidato está identificado por dos dígitos que corresponden al número aleatorio que se le asignó durante la prueba. Los ítems de la mitad derecha están identificados en función de la destreza evaluada (CA significa «comprensión auditiva»), de la tarea a la que pertenecen (esta prueba contenía tres tareas: A, B y C), y contienen tanto el número del ítem dentro de cada tarea como la respuesta correcta al mismo. Así pues, CAC7_C, en la parte superior derecha de la figura, es un ítem de comprensión auditiva que estaba contenido en la tarea C, dentro de la cual era el número 7 y cuya respuesta correcta era C. La forma de etiquetar tanto a candidatos como a ítems puede ser adaptada a distintas necesidades. La experiencia nos dice que, independientemente del formato, es útil tener identificada la respuesta correcta a un ítem en el nombre del mismo.

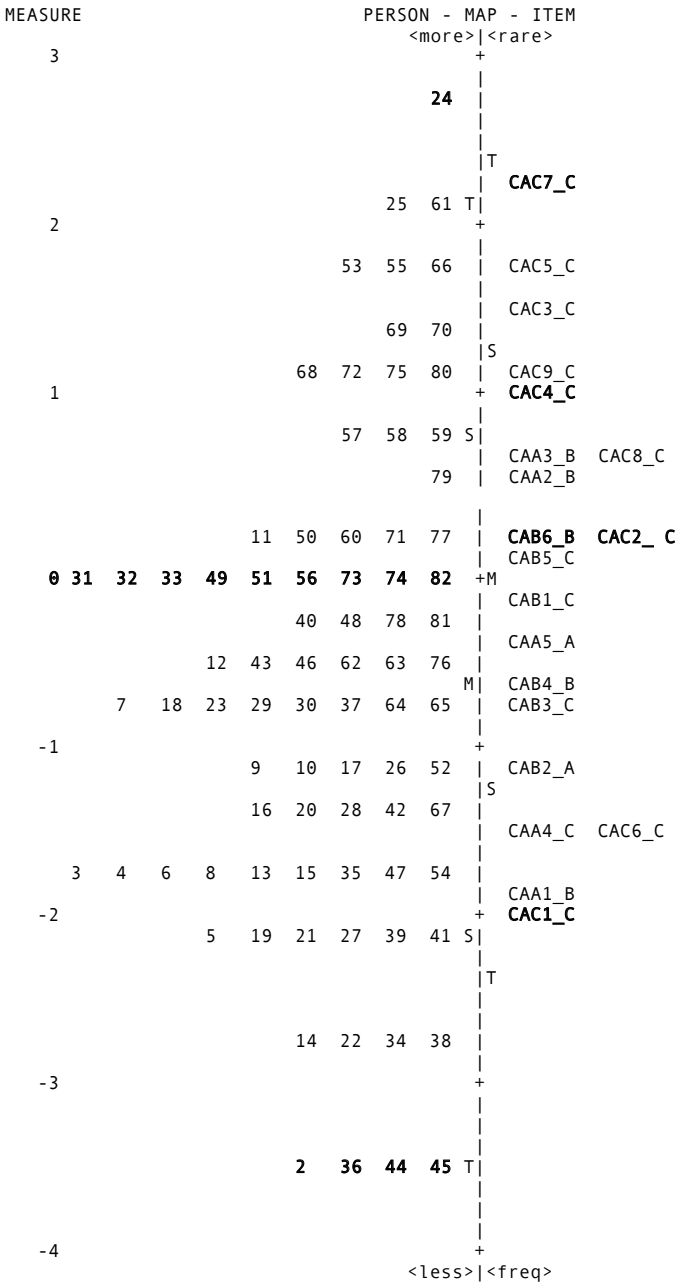


Figura 2. Mapa de la variable

Lo que esta figura nos dice es que el candidato más hábil es el candidato 24 (con una habilidad cercana a los tres *logits*) y que los candidatos menos hábiles son el 2, el 36, el 44 y el 45 (con una habilidad de en torno a los -3.5 *logits*). De igual manera nos indica que el ítem más difícil es el CAC7_C, con una dificultad de algo más de 2 *logits*, mientras que el más sencillo es el CAC1_C, que tiene una dificultad de alrededor de -2 *logits*.

Analizando la separación en *logits* entre ítems y personas en el mapa podremos calcular la probabilidad de acierto de los candidatos. En la tabla 7 se calcula esta probabilidad según la ecuación básica de Rasch (Wright, 1977):

Habilidad del candidato	Dificultad del ítem	Diferencia	Probabilidad
5	0	5	99 %
4	0	4	98 %
3	0	3	95 %
2	0	2	88 %
1	0	1	73 %
0	0	0	50 %
0	1	-1	27 %
0	2	-2	12 %
0	3	-3	5 %
0	4	-4	2 %
0	5	-5	1 %

Tabla 7. Probabilidad de acierto en función de la separación de *logits*

Como se observa, los rangos no ascienden de forma lineal, y a partir de ±3 *logits* de diferencia la probabilidad de acierto aumenta o disminuye dramáticamente. Si aplicamos la tabla 7 a la figura 2 observamos que cualquiera de los candidatos que aparecen alineados con el 0 en la figura 2 tendría una probabilidad del 27 % de responder correctamente al ítem CAC4_C dado que la diferencia en *logits* entre estos y el ítem es -1. De igual forma, cualquiera de estos mismos candidatos tendría una probabilidad de acierto en el ítem CAC1_C del 88 % puesto que la distancia que los separa es de 2 *logits*, etc. Estos cálculos se pueden utilizar en

cualquier otro rango y para mayor precisión se puede usar el valor exacto en *logits* de ítems y candidatos que se obtiene en tablas como la 6.

En la figura 2 también podemos apreciar la dispersión de los candidatos e ítems. Idealmente, nuestros ítems deberían estar distribuidos a lo largo de la variable latente que se muestra segmentada en *logits*. Esto significaría que hemos creado un conjunto de ítems con capacidad suficiente para medir distintos niveles de habilidad. Cuando varios ítems tienen el mismo valor en *logits* pueden estar indicando redundancia. Si se trata de redundancia métrica, esta no supone un problema (a más ítems, mayor precisión). Será la redundancia de contenido la que deba preocuparnos. En el caso de la figura 2, nunca hay más de 2 ítems con valores de *logits* similares, por lo que no parece haber redundancia de ningún tipo. De manera inversa, si existen segmentos del continuo de *logits* que no tienen ningún ítem asociado podemos estar ante puntos ciegos de nuestros exámenes. Además, si un conjunto de ítems está correctamente diseñado para una población específica no debería haber muchos candidatos por encima del nivel del ítem más difícil. Cuando esto ocurre significa que esos candidatos tienen un nivel mayor que el del ítem más difícil de la prueba. En el caso de la figura 2, por ejemplo, el candidato 24, el más hábil, tiene una habilidad superior a la del ítem CAC7_C, que es el más difícil.

En ocasiones es imposible diseñar ítems que cubran todo el continuo de habilidad de los candidatos que concurren al examen. En el contexto universitario español, por ejemplo, donde el nivel B1 en un idioma extranjero es requisito para terminar cualquier grado, es frecuente que candidatos con un nivel por encima o por debajo del B1 concurran a pruebas de este nivel para asegurar una calificación que les permita cerrar su expediente académico. En estos casos podemos encontrar lo que se conoce como «efecto techo» o «efecto suelo», que ocurre cuando nuestros ítems no pueden medir la habilidad de los candidatos más hábiles o menos hábiles con toda la precisión que sería deseable. En la figura 2 de más arriba encontrábamos un ejemplo de «efecto suelo». La precisión en las medidas de todas las personas que están por debajo del ítem CAC1_C (el más sencillo) es menor de lo que lo sería si tuviésemos ítems del nivel de estos candidatos. En este caso, el efecto suelo se explica porque el examen al que corresponde la figura 2 era gratuito, y es frecuente que a estos exámenes concurran personas que no están preparadas pero que desean probar suerte, algo que también puede ocurrir en las pruebas de acceso libre de las escuelas oficiales de idiomas. En la figura 3 encontramos un ejemplo de «efecto techo». La figura corresponde a una prueba realizada por la Unidad de Evaluación y Certificación en Lenguas de la Escuela de Idiomas de la Universidad de Antioquia, Colombia. Esta prueba multinivel estaba diseñada para medir los niveles B1, B2 y C1. Como vemos, a ella concurrieron candidatos (en este caso identificados con x) de nivel superior al C1, todos aque-

llos que están por encima de los ítems CL_24_II y CL_6_IIP. Es posible que esto ocurra en exámenes diseñados para medir una parte concreta del constructo.

Figura 3. Mapa de la variable con «efecto techo»

2.2. Tecnología y medición

Vivimos tiempos de avances sin precedentes en los que las palabras quedan desfasadas apenas se posan sobre el papel. Desde mediados del siglo XX somos testigos privilegiados de una revolución tecnológica comparable a la Revolución Industrial del XVIII. Sin duda, el progresivo aumento de la capacidad computacional del ser humano es uno de los principales facilitadores de estos avances. Este aumento se suele ilustrar mediante la ley de Moore, según la cual la capacidad de procesamiento de un ordenador se duplica cada dos años. Así se explica que un microchip del año 2015, comparado con otro de primera generación de 1971, tenga un rendimiento 3 500 veces superior, sea 90 000 veces más eficiente y cueste 60 000 veces menos. Si comparásemos desde esta misma óptica un Volkswagen Beetle de 1971 con otro de 2015, el de 2015 podría alcanzar los 480 000 km/h, consumiría tres litros de gasolina por cada 3 200 000 km y costaría tres céntimos (Del Jesus, 2021:23–24). La inteligencia artificial es solo la última manifestación de este vertiginoso crecimiento.

En las próximas dos secciones analizaremos, por un lado, algunas de las herramientas computacionales que nos permiten realizar los análisis psicométricos de los que hemos hablado y, por otro, cómo los avances tecnológicos en el campo de la computación pueden contribuir a la evaluación de lenguas.

2.2.1. Programas que convierten palabras en números

Afortunadamente, existen programas informáticos capaces de realizar los complejos y repetitivos cálculos que son necesarios para cualquier estudio psicométrico. Estos programas nos permiten centrarnos en los resultados sin necesidad de tener un conocimiento profundo del sustrato matemático de la psicometría. Al igual que no es necesario ser programador para usar *Google*, tampoco es necesario saber matemáticas para usar los programas que vamos a mencionar. Mencionaremos algunos de los que más tiempo llevan en uso y, por tanto, tal vez los más conocidos, sin que esto suponga necesariamente que los aquí referidos sean mejores que otros ya existentes o aún por desarrollar.

Probablemente el programa estadístico más popular desde finales del siglo XX sea el paquete *SPSS* de IBM (IBM Corporation, 2023). Este programa es especialmente recomendable para TCT y su principal ventaja radica en que, debido a que lleva varias décadas en uso, existen multitud de herramientas de apoyo para su empleo (*vid.* Field, 2014). Una simple búsqueda en internet es suficiente para encontrar un sinfín de tutoriales que suelen ser de gran ayuda. Otra ventaja de *SPSS* de IBM (IBM Corporation, 2023) es el hecho de que su interfaz ha sido pulida a lo largo de los años y, aunque no llega a ser tan intuitiva como la de un editor de texto, se asemeja bastante a la de las hojas de cálculo con las que la mayoría estamos familiarizados. *SPSS* (*ibid.*) es una potente herramienta que po-

sibilita realizar multitud de análisis, exportar tablas y gráficas de forma sencilla e incluso llevar un registro de la evolución de los diferentes análisis realizados que nos permite volver sobre nuestros pasos en cualquier momento.

En el ámbito de la otra tradición psicométrica analizada en este libro, a principios de la década de 1970 «empiezan a aparecer los programas informáticos necesarios para utilizar los modelos de TRI, tales como BICAL y LOGIST en 1976, BILOG en 1984, MULTILOG en 1983 y otros muchos» (Muñiz, 2010:63). De entre todos estos, el programa que ha gozado de mayor popularidad desde su lanzamiento en 1987 es *Winsteps* (Linacre, 2024b). Todos los análisis de la sección 2.1.2 están realizados con *Winsteps* (*ibid.*). Las principales ventajas de este programa de carácter minimalista, pensado fundamentalmente para análisis Rasch, son la gran cantidad de publicaciones que pueden encontrarse en internet sobre su uso y el manual digital de ayuda que incorpora, que permite aprender a usar el propio paquete informático e investigar sobre los fundamentos matemáticos del modelo Rasch. *Winsteps* (*ibid.*) es a la vez una potente herramienta informática y el libro más completo que jamás se ha escrito sobre el modelo Rasch. En la parte negativa mencionaremos su apariencia poco atractiva, poco intuitiva, y el hecho de que, al no tener una gran corporación tras de sí, no está claro si se podrá seguir usando cuando los actuales desarrolladores dejen de ofrecer soporte.

Facets (Linacre, 2024a) es el hermano gemelo de *Winsteps* (Linacre, 2024b). Mientras que *Winsteps* (*ibid.*) se usa para analizar datos dicotómicos, *Facets* (Linacre, 2024a) se usa para análisis politómicos en los que pueden concurrir diferentes facetas (candidatos, ítems, evaluadores, intentos, etc.), por lo que está particularmente indicado para el análisis de escalas de evaluación. En lo práctico y lo estético comparte las mismas virtudes y carencias de *Winsteps* (Linacre, 2024b), es decir, un potente manual de ayuda y una interfaz poco intuitiva. Todos los análisis del capítulo 4 están realizados con *Facets* (Linacre, 2024a).

Una alternativa interesante a todos estos programas es *R* (R Development Corte Team (2024). La principal ventaja de este programa es que, al estar diseñado con código abierto, es gratuito y cualquier usuario puede crear nuevos paquetes de análisis, por lo que su continuidad está garantizada en tanto que exista apoyo de su comunidad de usuarios. Esto lo hace también adaptable a nuevas formas de análisis. *R* (*ibid.*) es sin lugar a dudas el menos intuitivo de todos los programas de los que hemos hablado hasta ahora. Está compuesto por una primera capa cuyo uso requiere ciertos conocimientos de programación. A través de esta primera capa se ejecutan los diferentes paquetes de análisis que tienen como principales ventajas la gran calidad de los gráficos generados y la versatilidad de los paquetes ejecutables.

Además de estos programas cabe mencionar otros como *Jamovi* (The Jamovi Project, 2024), *MATLAB* (The MathWorks Incorporated, 2022), *Mathematica* (Wolfram Research Incorporated, 2024) y *XCalibre* (Assessment Systems Corporation,

2014), etc. o incluso *Excel* (Microsoft Corporation, 2024), que a través del uso de fórmulas también permite el análisis estadístico. La elección de uno u otro dependerá del contexto, tradición y orientación del investigador.

2.2.2. Tecnología y futuro

Más allá de los programas informáticos que nos permiten trabajar con la psicometría, avances tecnológicos como los que mencionábamos en el capítulo 1 modificarán, sin duda, las necesidades de los futuros estudiantes y candidatos a exámenes de dominio de lengua.

No obstante, al igual que la invención de la luz eléctrica no nos ha hecho renunciar a la luz del sol, no hemos de pensar que las redes neuronales o la inteligencia artificial nos harán abandonar el aprendizaje, la enseñanza y la evaluación de lenguas. Así como la luz eléctrica nos permite hacer cosas impensables antes de su invención, la creciente capacidad computacional del ser humano nos conducirá a lugares que hoy son difíciles de imaginar.

Hoy ya podemos concebir un futuro próximo en el que un candidato realice un examen de dominio en línea vigilado por un sistema de seguimiento ocular (*vid*. 3.4.3) en el que su capacidad de comprensión auditiva y lectora se evalúen mediante algoritmos adaptativos implementados en una inteligencia artificial que interactúa oralmente y evalúa corrección y riqueza léxica. En este escenario, la intervención humana se vería reducida notablemente, tal vez quedando relegada a la revisión de las estimaciones de las inteligencias artificiales en casos dudosos y a la corrección de aquellos aspectos de la producción que dichas inteligencias no están aún en condiciones de evaluar. Todos estos cambios han de conducirnos a diferentes reflexiones sobre el constructo de nuestras pruebas. Hemos de plantearnos si las competencias tecnológicas de un candidato afectan o no a su desempeño en las pruebas igual que lo haría (o no) su capacidad de escribir o leer más o menos rápido. Asimismo, hemos de plantearnos si estamos aceptando todos estos cambios desde el convencimiento de que mejoran la validez de nuestras pruebas o, sencillamente, porque son los signos de los tiempos.

CAPÍTULO 3

Los Exámenes

Se espera que los evaluadores de lenguas den respuesta y sean sensibles a teorías derivadas de dos ámbitos tan alejados entre sí como lo son la lingüística (que describe el conocimiento del lenguaje) y la psicometría (que mide dicho conocimiento y otros atributos humanos), y se espera además que lo hagan dentro de los límites impuestos por factores institucionales, económicos, sociales e incluso políticos. (Spolsky, 1995:4)

Los exámenes no son, ni mucho menos, una invención moderna. Ya en la *Biblia* se nos habla del que tal vez sea el primer examen de pronunciación del que tenemos constancia. Este podría haber tenido lugar en el siglo XII a. C., época en la que vivió Jefté, uno de los jueces de Israel:

> Entonces Jefté reunió a todos los hombres de Galaad y atacó a Efraím. Los de Galaad derrotaron a los efraimitas, que decían despectivamente: «Vosotros, los de Galaad, sois fugitivos de Efraím, en medio de Manasés». Galaad ocupó los vados del Jordán para cortarle el paso a los Efraimitas. Y cuando un fugitivo de Efraím intentaba pasar, los hombres de Galaad le preguntaban: «¿Tú eres de Efraím?» Si respondía que no, le obligaban a pronunciar la palabra *Shibolet*, pero decía *Sibolet*, porque no podía pronunciar correctamente. Entonces lo prendían y lo degollaban junto a los vados del Jordán. En aquella ocasión murieron cuarenta y dos mil hombres de Efraím. (Jueces, 12:4–6)

Por descabellado que parezca, se conocen casos similares al de *shibolet* (שִׁבֹּלֶת en hebreo) en fechas recientes. Durante los primeros compases de la invasión rusa en Ucrania, en 2022, muchos soldados rusos se vistieron de civiles para infiltrarse en las líneas ucranianas aprovechando el hecho de que la mayoría de la población ucraniana habla ruso y ucraniano, dos lenguas que comparten muchas características. En primera instancia, cuando los ucranianos fueron conscientes de ello, comenzaron a destruir las señales que contenían nombres de ciudades y calles para entorpecer las incursiones enemigas. De esta manera forzaban a los espías rusos a preguntar por indicaciones en los puntos de control, algo que podía dejarlos al descubierto en caso de verse obligados a hablar en ucraniano. Con el tiempo, y conforme los soldados rusos comenzaron a usar también el ucraniano en sus incursiones, los habitantes locales tuvieron la idea de usar la palabra паляни́ця, cuya transliteración es *palianytsia*, que hace referencia a un tipo de pan tradicional ucraniano. Esta palabra, de uso infrecuente en la vida cotidiana,

cuyas consonantes palatales son de difícil pronunciación para los hablantes no nativos, pronto se convirtió en el *shibolet* ucraniano.

Uno de los autores que más y mejor nos ha hablado sobre la historia de los exámenes, Spolsky (1977; 1995; 2017), considera las pruebas imperiales chinas centralizadas del siglo XII como el primer intento de establecer un sistema de evaluación estandarizado. El objeto de estos exámenes era el de escoger a los mejores candidatos para el puesto de mandarines, los altos funcionarios de la china imperial.

El sistema centralizado chino sería exportado a Europa por los jesuitas en el siglo XVI. A su llegada a Europa, el método chino se modifica y pasa a ser usado para evaluar el avance de los alumnos de las escuelas cristianas europeas hasta el siglo XVIII. De esta manera, se produce un cambio en el objetivo último de los exámenes, que pasan de utilizarse para elegir al mejor de entre un grupo de candidatos, a convertirse en una herramienta para seguir el avance de los estudiantes. Tras la Revolución Francesa, Napoleón implantó los exámenes al final de la educación secundaria, un sistema que se popularizó en muchos otros países, donde aún hoy se mantienen (el *Abitur* en Alemania; los *A-Level* en Reino Unido; la EVAU/PAU en España; el ENES en Ecuador; el EXANI-II en México; el *Matura* en Austria; el Saber 11º en Colombia; el PSU en Chile; los SAT en Estados Unidos, etc.).

A comienzos del siglo XX, con la eclosión de la psicometría (*vid.* 2.1), se popularizarán en Estados Unidos los tests de inteligencia, cuyas mecánicas fueron usadas con múltiples fines (Spolsky, 2017:376-377), algunos de ellos espurios. A partir de ese momento, la proliferación de exámenes cuyo principal objetivo es la estandarización de los candidatos ha llegado a condicionar la enseñanza (un efecto que en inglés se conoce como *washback*), por lo que la relación entre esta y aquellos ha de ser cuestionada y revisada constantemente. En la actualidad, por ejemplo, tienen cada vez menos sentido los métodos de evaluación generales basados en la memorización que han primado después de la Revolución Industrial, o los métodos de evaluación derivados de la enseñanza de lenguas muertas, que obvian por completo la interacción y la mediación. Hoy día parece más deseable que nuestras pruebas evalúen las competencias necesarias para desenvolverse en contextos pluriculturales y plurilingües en los que el uso de herramientas digitales sustituye a la memorización. Si nuestros exámenes han de reflejar las tareas de la vida diaria, parece lógico que el uso de herramientas como internet o la inteligencia artificial pasen a ser parte de nuestros exámenes.

A mediados del siglo XX, impulsados en primera instancia por las necesidades derivadas de la Segunda Guerra Mundial (Spolsky, 1995:99–116), los exámenes se industrializaron y comenzaron a distribuirse a gran escala a partir de la década de 1960. La psicometría dio una base racionalista y empírica a los argumentos sobre la fiabilidad de las pruebas. Acto seguido comenzó la reflexión sobre la compleja

cuestión de la validez. Habiendo superado el debate acerca de la necesidad de fiabilidad y estabilidad en las mediciones, había que interpretar el significado de estas últimas. Es en este momento cuando aparecen las reflexiones de Cronbach y Meehl (1955), de Messik (1989), Bachman (1991) y, más recientemente, Weir (2005), por citar solo algunos nombres. Como veremos en la sección 3.3, la idea de validez se atomiza en esta época y se vuelve casi impracticable hasta la llegada de Messick, quien propone considerar la validez no como un concepto unitario, sino como un prisma de múltiples caras que se construye a través de evidencias de distinta procedencia. Según esta visión, que nosotros compartimos, no existe un único guarismo, una única cifra o análisis estadístico que determine si un examen es válido o no. Estas cifras, junto con otras pruebas (el criterio de expertos, la vinculación de nuestros exámenes a un estándar, la equidad en la administración de las pruebas, etc.), son las aristas que unen las distintas caras del complejo prisma que es la validez.

La importancia de los exámenes en la sociedad actual es tal que estos se filtran en todas las capas de nuestra cultura. En su novela *La fiesta del Chivo* (Vargas, 2000:221), Vargas Llosa recrea una cena entre el dictador dominicano Rafael Leónidas Trujillo y sus allegados militares que habría ocurrido en algún momento de la primera mitad del siglo XX, época en la que Trujillo gobernó. En la ficción de Vargas Llosa (que describe un episodio similar al del *shibolet* judío y el *palianytsia* ucraniano), Simon Gittleman, mentor de Trujillo, le pregunta a este sobre la Matanza del Perejil:

> —¿Es verdad lo del perejil, Su Excelencia? ¿Que para distinguir a dominicanos de haitianos se hacía decir a los negros *perejil*? ¿Y que a los que no la pronunciaban bien les cortaban la cabeza?
> —He oído esa anécdota —se encogió de hombros Trujillo—. Habladurías que corren por ahí.

El pasaje hace referencia a uno de los hechos más dramáticos acaecidos durante la dictadura de Trujillo, la Masacre del Perejil, en la que decenas de miles de haitianos murieron a manos de soldados dominicanos. Los países de Haití y la República Dominicana comparten isla y fronteras, por lo que es habitual el tránsito de emigrantes de un país a otro, principalmente del primero al segundo. Los haitianos, de rasgos afroamericanos y practicantes de vudú, eran vistos como una amenaza por la dictadura de Trujillo. En Haití se habla criollo, una lengua de origen francés con influencias africanas. Las características fonéticas del criollo haitiano y del español de la República Dominicana son, pues, distintas. Se cree que, sobre la base de las diferencias fonéticas entre ambas lenguas, las tropas de Trujillo asesinaron a muchos de los 20 000 haitianos muertos durante la Masacre del Perejil. En octubre de 1937 los soldados del Chivo (apodo por el que era co-

nocido el dictador) pidieron de forma sistemática a los sospechosos de ser haitianos residentes en la República Dominicana que pronunciasen la palabra «perejil». Obviamente, los hablantes del criollo haitiano lo hacían de una forma distinta a los hablantes dominicanos del español. De esta manera, con la pronunciación de una palabra, miles de haitianos firmaron su sentencia de muerte.

Otro ejemplo es el del que, sin duda, se ha convertido en el examen más famoso de la historia de la literatura y del cine, la prueba Voight-Kampf:

> Que él recordara, cerca de cincuenta androides modelo T-14 lograron de un modo u otro colarse en la Tierra, y en ciertos casos se tardó hasta un año en detectarlos. Pero entonces se diseñó el llamado test de empatía Voight, obra del Instituto Pavlov de la Unión Soviética. Al menos que él supiera, ningún androide T-14 había logrado burlar ese test. (Dick, [1968] 1993:26–27)

Este pasaje de la novela *Do Androids Dream of Electric Sheep* (Dick, [1968] 1993), que inspiró la película *Blade Runner* (Scott, 1982), describe el sueño de cualquier redactor de exámenes: la prueba perfecta e infalible que sirve para diferenciar a quienes poseen una característica de quienes no la tienen, sea esta la de hablar un idioma o la de ser un replicante.

En el lado opuesto a la objetividad de la prueba Voight-Kampf, tenemos sorprendentes ejemplos de exámenes cuya subjetividad está fuera de toda duda. En el ámbito de las artes marciales japonesas, por ejemplo, encontramos el curioso caso de los exámenes de 8º *dan* de *kendo*. El *kendo* es un arte marcial surgido en Japón tras la restauración Meiji de 1868, cuyos practicantes son fácilmente reconocibles por el uso de armadura y sables de bambú. El grado máximo que se puede ostentar en esta disciplina es el de 8º *dan*, y para alcanzarlo pueden ser necesarios más de cincuenta años de dedicada práctica. Los exámenes de 8º *dan* están considerados como unos de los más difíciles del mundo en cualquier disciplina. El porcentaje de aprobados en estos exámenes varía según las convocatorias y las fuentes que se consulten, si bien suele estar por debajo del 1 % de entre los 700 a 1000 candidatos que concurren. Aunque la estructura de estos exámenes ha variado a lo largo de los años, una parte central de ellos siempre ha sido el combate directo entre dos espadachines rivales. Para lograr el éxito en el combate, ambos oponentes buscarán golpearse con un *yuukoo datotsu* (有効打突 en japonés) es decir, con una estocada correcta realizada con espíritu elevado, con la postura adecuada, usando el tercio superior del sable e impactando con el ángulo justo en las partes designadas de la armadura del oponente (International Kendo Federation, 2017:5). Esta estocada, extremadamente rápida, ha de estar acompañada por un grito (o *kiai*) y a su conclusión el *kendoka* habrá de mostrarse en un estado de alerta física y mental (o *zanshin*). Solo cuando confluyen estos factores tiene lugar la correcta unidad entre la espada, el cuerpo y el espíritu (o *kikentai no itchi,* 気剣体の一致 en japonés)

que conduce al golpe correcto (Bennett, 2015:xxx–xxxi). Como se aprecia, la definición de lo que es correcto o no en estas pruebas incorpora aspectos de difícil observación directa (espíritu elevado, alerta mental) que, a su vez, están sujetos al escrutinio cuasi-esotérico de los hasta catorce jueces que pueden llegar a intervenir en una prueba, y que interpretarán la normativa según los criterios que les dicte su experiencia individual, escrutinio que nunca es cuestionado por los aspirantes. El espíritu de superación y disciplina que rodea al *kendo* es admirable, si bien se podría convenir que, desde la perspectiva del observador externo, la mezcla de criterios objetivos y otros que no lo son tanto puede suponer una fuente de frustración en exámenes de este tipo.

Dejando al margen los referentes históricos, literarios y marciales, actualmente estamos viviendo tiempos de profundos cambios en la evaluación en general y en la evaluación de lenguas en particular. La tecnología ha penetrado en nuestro campo en las últimas décadas, espoleada, entre otras cosas, por una pandemia mundial. Ya existen multitud de exámenes de dominio que se realizan exclusivamente en línea, algunos de los cuales permiten evaluar a los candidatos desde su propio hogar. Más allá de las preocupaciones de seguridad que envuelven a este formato de exámenes, existe una reflexión que afecta a nuestros constructos. La evaluación de lenguas es una parte más del proceso de experiencia social que genera el aprendizaje de lenguas. Hemos de reflexionar sobre el objetivo de nuestra evaluación y sobre la forma en que se presentan nuestros exámenes, por un lado, y sobre cómo todo lo anterior modifica nuestro constructo al tiempo que refleja el espíritu de la época en que tienen lugar.

El objetivo de nuestra evaluación y la forma en que esta se presenta están íntimamente ligados. La forma de comunicación en redes sociales, por ejemplo, es ostensiblemente diferente de la forma de comunicación que tiene lugar cara a cara. Nos guste o no, si esta forma de comunicación se instala como algo habitual en nuestras sociedades, habremos de estar preparados para evaluarla también en nuestras pruebas. El debate está servido, y las implicaciones, por investigar. En ocasiones el objetivo puede condicionar la forma y viceversa. La forma en que diseñamos y presentamos nuestras pruebas ya está siendo transformada por los avances tecnológicos. Los algoritmos de las pruebas adaptativas multinivel, la evaluación semi-directa de producción oral y escrita mediante dispositivos electrónicos, la corrección automatizada de tareas de producción oral y escrita o la evaluación de producción oral mediante inteligencia artificial son solo algunos ejemplos que parecían impensables a principios del siglo XXI y que hoy deberían estar en el centro del debate sobre la ética de nuestras pruebas. Es imperativo reflexionar sobre la manera en que la tecnología se integra en nuestros exámenes. No es suficiente con el deseo de utilizarla: es necesario conocerla y prepararse para usarla. La forma en que usemos la tecnología debe ser un reflejo de la sociedad de la que emana y afectará a los resultados obtenidos a través de ella, que, a su vez, tendrán

efectos sobre la comunidad de evaluadores y dará forma a nuestra disciplina en un círculo que debería ser virtuoso, no vicioso. Las inteligencias artificiales y sus algoritmos, no obstante, no sustituirán a los evaluadores, no al menos en el siglo XXI, pero ocuparán un lugar cada vez más relevante.

Con todo esto en mente, como indicábamos en la sección 2.2.2, no es difícil imaginar un futuro próximo en el que un candidato realice un examen de dominio en línea vigilado por un sistema de seguimiento ocular (*vid.* 3.4.3) en el que su capacidad de comprensión auditiva y lectora se evalúen mediante algoritmos adaptativos implementados en una inteligencia artificial que interactúa oralmente y evalúa el grado de corrección y riqueza léxica de un candidato. Todos estos cambios han de conducirnos a diferentes reflexiones sobre el constructo de nuestras pruebas.

Habiendo sugerido algunas líneas de pensamiento sobre el futuro de la evaluación de lenguas, adoptaremos a continuación un enfoque más práctico y descriptivo. Para ello, en este capítulo primero profundizaremos en la definición de los exámenes. En segundo lugar, y dado que ningún examen tiene lugar en el vacío, analizaremos lo que ha supuesto el MCER (Consejo de Europa, 2001) para la evaluación de lenguas. A continuación dedicaremos una sección a reflexionar sobre el argumento de validez de nuestras pruebas y, por último, trataremos de dar pautas para el diseño de los diferentes componentes de nuestros exámenes. Tal vez, en un futuro no muy lejano, estas pautas se vean tan primitivas como hoy se ven otras ya pasadas pero que, en cualquier caso, presentan las cuestiones pertinentes de forma racional y ordenada. Si todos los avances científicos descansan sobre los esfuerzos del pasado, esperamos que las reflexiones que ahora siguen contribuyan al avance de nuestra disciplina en el futuro.

3.1. Evaluación, medición y exámenes

Para diferenciar entre estos conceptos es necesario caracterizarlos. La medición ya quedó definida en el capítulo 2, donde vimos que consiste en comparar una cantidad indeterminada de algo con su respectiva unidad, bien sea esta el kg (al pesar una bolsa de naranjas), los *logits* (al considerar la dificultad de un examen de dominio de lengua), o cualquier otra.

Evaluar, por otro lado, es recoger datos de forma sistemática con el objeto de tomar decisiones (Bachman, 1991:22), algo que no tiene por qué ir de la mano de los exámenes (pensemos, por ejemplo, en las anotaciones de un profesor sobre el avance de un alumno). Un examen, por otro lado, sería una herramienta de medida creada específicamente para obtener una muestra concreta del comportamiento de un candidato (por ejemplo, cómo este interactúa verbalmente). Así pues, «evaluar» y «examinar» no son sinónimos. El diagrama de interacciones entre estos términos que Bachman (*ibid.*:23) propone es el siguiente:

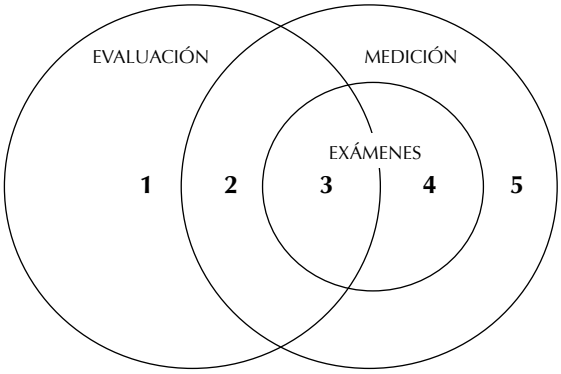


Figura 4. Relación entre evaluación, medición y exámenes

En el área 1 de su diagrama, Bachman ubica la evaluación que no conlleva mediciones ni exámenes, por ejemplo, las anotaciones de un profesor sobre el avance de un alumno de las que hablábamos anteriormente. Bachman no da un ejemplo claro para el área 2. Nosotros proponemos imaginar una clase en la que el profesor otorga puntos a sus estudiantes por determinadas acciones del día, puntos que se sumarán al final del curso para premiar a los alumnos más colaboradores. Los exámenes de aprovechamiento usados para comprobar el progreso de un alumno al final de una unidad estarían en el área 3, mientras que los exámenes de dominio de lengua que son el eje principal de este libro estarían en el área 4. Por último, un ejemplo para el área 5 sería el del lingüista que asigna a los diferentes participantes en un estudio un número en función de las lenguas que hablan (1 para el participante que habla una lengua, 2 para el que habla dos, etc.).

Tan importante es distinguir entre estos tres conceptos como, ya centrados en la evaluación, conocer las distintas formas en que esta puede presentarse. El MCER (Consejo de Europa, 2001:183), por ejemplo, establece hasta veintiséis tipos de evaluación diferentes.

Esta diferenciación de formatos de evaluación está incluida en el capítulo 9 del MCER (Consejo de Europa, 2001), cuya lectura es aconsejable si se desea profundizar en la conceptualización de la evaluación. Dentro de dicho capítulo también se definen los distintos tipos de evaluación de la tabla 8, que pueden darse por separado o aparecer juntos. Pensemos, por ejemplo, en un examen de producción oral dentro de una prueba de dominio en el que los candidatos han de interactuar conversando sobre un tema concreto mientras uno o más examinadores les aplican unas escalas de evaluación. En tal caso estaríamos ante un caso de evaluación de dominio, directa, objetiva en la que se realiza una evaluación analítica mediante una escala.

Evaluación de aprovechamiento	Evaluación de dominio
Con referencia a la norma	Con referencia al criterio
Maestría	*Continuum*
Evaluación continua	Evaluación en un momento concreto
Evaluación formativa	Evaluación sumativa
Evaluación directa	Evaluación indirecta
Evaluación de la actuación	Evaluación de los conocimientos
Evaluación subjetiva	Evaluación objetiva
Valoración mediante lista de control	Valoración mediante escala
Impresión	Valoración guiada
Evaluación global	Evaluación analítica
Evaluación en serie	Evaluación por categorías
Evaluación realizada por otras personas	Autoevaluación

Tabla 8. Diferentes tipos de evaluación según el MCER

Tener una idea clara de las diferencias que hay entre evaluación, medición y exámenes, y conocer las distintas formas que la evaluación puede adoptar es de vital importancia. Y lo es desde un punto de vista práctico, por ejemplo, para no diseñar un examen de dominio como si se tratase de un examen de aprovechamiento. En un examen de dominio mediremos con respecto a un estándar externo (los niveles del MCER (Consejo de Europa, 2001), por ejemplo), mientras que un examen de aprovechamiento mediremos con respecto a los contenidos del curso. Los candidatos a un examen de dominio pueden haberse formado en diferentes contextos y tener poco en común entre sí, mientras que los candidatos a un examen de aprovechamiento suelen proceder de un contexto educativo similar y comparten un mismo programa de estudios.

Esta primera sección del capítulo dedicado a los exámenes nos ha conducido a centrar la atención en un tipo concreto de herramienta de medición: las pruebas que diseñaremos para obtener respuestas lingüísticas específicas de nuestros candidatos y que llamamos exámenes de dominio (frente a, por ejemplo, los exámenes de progreso). Definida nuestra herramienta, en las próximas secciones analizaremos primero contexto en el que esta se usa, la noción de validez y, más tarde, sus componentes principales.

3.2. Europa, el *Marco Común Europeo de Referencia* (MCER) y el *Volumen Complementario* (MCERVC)

Ningún examen tiene lugar en el vacío. Conocer el contexto de la evaluación de lenguas en el siglo XXI en Europa y en gran parte del mundo implica conocer el Viejo Continente y el impacto que en él han tenido dos obras seminales. Estas obras han tenido un impacto tal que no solo se han convertido en el marco de referencia en Europa sino que, además, han dado origen a versiones adaptadas en otras partes del mundo. Como el título de esta sección indica, nos estamos refiriendo al MCER (Consejo de Europa, 2001) y al MCERVC (Consejo de Europa, 2020).

Para entender el MCER (Consejo de Europa, 2001) se ha de entender la idea de Europa, una idea que es esquiva a las afiliaciones políticas y a los límites geográficos. Desde hace siglos los europeos hemos intentado capturar nuestra realidad de diferentes maneras. Una de ellas ha sido la creación de la Unión Europea, formada actualmente por veintisiete países que comparten objetivos económicos, políticos y un sentimiento histórico de hermandad.

Pero como decíamos anteriormente, la idea de Europa es esquiva a los límites políticos y también a los geográficos. A algunos lectores tal vez les resulte sorprendente saber que países como Reino Unido, Islandia, Noruega o Suiza no son parte de la Unión Europea y que otros como Finlandia sí lo son. De hecho, es tan fácil encontrar a ciudadanos ingleses, islandeses, noruegos o suizos que se consideran europeos como difícil es no sentirse europeo dentro de estos países. Términos como Eurasia (relacionado con la geología y con la distopía *1984* de Orwell (1950)), Zona Euro (que designa a los veinte países que comparten la moneda única europea) o Espacio Schengen (el conjunto de países entre los que los europeos podemos circular sin pasaporte) son un reflejo de la riqueza y de la complejidad cultural, social y política del Viejo Continente.

Al margen de los dictados de la geopolítica, los países que de una manera u otra dialogan en el ágora europea lo hacen en muy diversas lenguas. Esto aporta una indudable riqueza cultural y lingüística a Europa que es, a la misma vez, un tesoro y un inconveniente en tanto que los europeos no disponemos de una lengua común en la que poder comunicarnos.

Desde hace más de mil años, antes incluso de que la propia idea de Europa comenzarse a fraguarse, gran parte de esta diversidad cultural y lingüística ha sido articulada en torno a las universidades. Las universidades europeas han resistido el paso del tiempo, la política, la religión y la guerra, manteniéndose constantes como agentes culturales, guardianes del pasado y brújulas para el futuro. Conscientes de su importancia, en 1988, 388 universidades europeas firmaron la *Magna Charta Universitatum* (MCU, 1988) durante la celebración del noveno centenario de la fundación de la Universidad de Bolonia, la más antigua de Europa. Este

documento sentó las bases modernas de la libertad académica, de la autonomía institucional y ha guiado desde entonces a las universidades hacia el mutuo entendimiento. Fruto de este documento y del descrito deseo de entendimiento se redactó en 1999 la *Declaración de Bolonia* (1999), que dio origen al conocido como Proceso de Bolonia. El Proceso de Bolonia, que pretende dar mayor coherencia a los sistemas de educación superior en Europa, estableció el Espacio Europeo de Educación Superior para facilitar la movilidad de estudiantes y de personal, para hacer que la educación superior fuese más inclusiva y accesible, y para hacer más atractiva y competitiva la educación superior en Europa. Con objeto de conseguir la referida coherencia 1) se introdujo un sistema de educación superior de tres ciclos (grado, máster y doctorado), 2) se garantizó el reconocimiento mutuo de las cualificaciones y los periodos de aprendizaje completados en universidades pertenecientes a los países firmantes y 3) se implementó un sistema de garantía de la calidad del aprendizaje y la enseñanza. Los movimientos de convergencia y mutuo reconocimiento impulsados por el Proceso de Bolonia han facilitado la movilidad internacional de varias generaciones de estudiantes europeos que, gracias principalmente a los programas Erasmus, han vertebrado de forma invisible pero robusta la tan perseguida idea de Europa. Quienes hemos tenido la suerte de participar en algún programa Erasmus percibimos la posibilidad de haber podido estudiar en otros países europeos como una vivencia transformadora y animamos a nuestros hermanos hispanoamericanos a replicar esta experiencia de éxito. Sin duda, sería tremendamente enriquecedor contar con un programa panamericano similar al europeo.

En paralelo a la construcción de la noción de Europa y tras dos guerras mundiales, en 1949 se creó el Consejo de Europa, que no se ha de confundir con el Consejo Europeo ni con el Consejo de la Unión Europea. El Consejo de Europa, que trasciende a la Unión Europea, está formado por cuarenta y seis Estados que buscan una unión más estrecha entre sí con el objeto de promover los ideales y los principios que constituyen su patrimonio común y favorecer su progreso económico y social (Consejo de Europa, 1949) así como la democracia, los derechos humanos y el Estado de derecho en Europa y en el mundo. El Consejo de Europa es precisamente el impulsor de los dos libros más influyentes en la historia europea de la enseñanza de idiomas que mencionábamos anteriormente: el MCER (Consejo de Europa, 2001) y el MCERVC (Consejo de Europa, 2020).

En la década de 1990, Brian North, el académico cuyo trabajo iniciaría la redacción del MCER (Consejo de Europa, 2001), del que es coautor, se encontraba trabajando una tesis doctoral que formaba parte de un proyecto de investigación suizo vinculado a la iniciativa del Consejo de Europa de crear el MCER (*ibid.*). El objetivo de este trabajo doctoral era crear el prototipo de un «Pasaporte de

Lenguas» que reflejase las competencias del alumnado en diferentes etapas de aprendizaje. La idea era hacer transparentes los difusos criterios que determinados colectivos de profesores suizos habían internalizado para convertirlos en definiciones claras de distintos niveles de desempeño. Se buscó pasar de criterios intuitivos a los criterios objetivos que *a posteriori* se materializarían en los conocidos *can-do statements* (i.e. las definiciones de lo que los usuarios de una lengua son capaces de hacer en determinados puntos del continuo de su aprendizaje) (North, 2000:1). Ante la constatación de que la mayoría de las definiciones de nivel de dominio existentes en aquel entonces parecían estar diseñadas desde la intuición o desde la valoración subjetiva de muestras de competencia (*ibid.*:3), el Consejo de Europa decidió desarrollar un marco científico de referencia con el objetivo de 1) ayudar a los distintos agentes involucrados en la enseñanza y el aprendizaje de idiomas a planificar el desarrollo de las capacidades necesarias para un uso efectivo del lenguaje en diferentes etapas; 2) capacitar a diferentes proveedores de servicios de enseñanza a describir y comparar sus sistemas mediante un metalenguaje común, y 3) aportar descripciones escalables de niveles de dominio lingüístico en diferentes categorías (*ibid.*:1–2). Todo este trabajo de desarrollo quedó reflejado en diversas publicaciones (North, 1995; 2000; North y Schneider, 1998) y culminó en el celebérrimo MCER (Consejo de Europa, 2001).

La versión final del MCER (Consejo de Europa, 2001) se publicó en 2001 y ha sido traducida a más de cuarenta lenguas (Consejo de Europa, 2020:21). El MCER (Consejo de Europa, 2001) es un documento consciente del entorno en el que nace y dedica su primer capítulo a describir el contexto político del que emana. El MCER (*ibid.*) se reivindica a sí mismo, sin inclinarse por un marco teórico concreto, como la base para la elaboración de programas de enseñanza, exámenes y libros de texto en Europa a partir del mismo momento de su publicación. El MCER (*ibid.*) nació porque los distintos movimientos que condujeron a la creación de la Unión Europea habían evidenciado, más que ninguna otra aproximación teórica, la necesidad de definir objetivos lingüísticos comunes (*vid.* Spolsky, 1995:3). No obstante, el verdadero éxito del MCER (Consejo de Europa, 2001) es haber dotado de un lenguaje común a los distintos agentes involucrados en la enseñanza, aprendizaje y evaluación de lenguas. Si este lenguaje común no hubiese sido útil habría sido abandonado inmediatamente y, sin embargo, décadas después persiste e incluso se ha intentado replicar en distintas partes del mundo. Como vemos, en el MCER (*ibid.*) cristalizan los ideales de la Unión Europea, el deseo de entendimiento mutuo de sus instituciones académicas, el reconocimiento explícito de la variedad lingüística y cultural del continente, y el poder catalizador del Consejo de Europa.

Dentro del lenguaje común que el MCER (Consejo de Europa, 2001) nos propone, sin duda, lo que mayor repercusión ha tenido es la estructura de niveles de dominio que este describe:

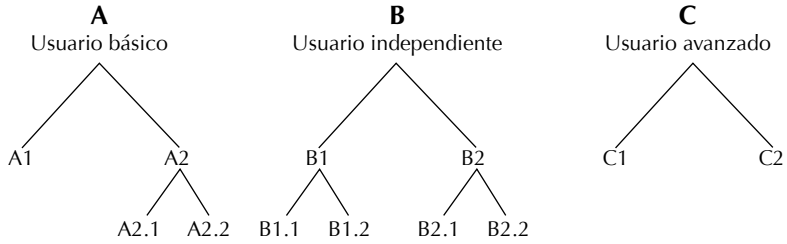


Figura 5. Estructura de niveles de dominio propuesta por el MCER

A esta estructura inicial se añadirían posteriormente niveles previos al A1 (Consejo de Europa, 2020:22) e incluso se mencionaría la existencia de un nivel de competencia más allá del C2, el denominado ambilingualismo (Consejo de Europa, 2020:37), todo lo cual se articula en tablas que ordenan los *can-do statements* antes mencionados que definen estos niveles.

El MCER (Consejo de Europa, 2001) utiliza, además, las ideas centrales de competencias (generales y específicas), estrategias y tareas. Simplificando mucho, podríamos decir que las competencias constituyen el conjunto de saberes que los usuarios de una lengua acumulan (North, 2000:41–54) y que activan mediante determinadas estrategias (Consejo de Europa, 2001:57–90) para llevar a cabo tareas de la vida diaria (*ibid*.:157–158). Las estrategias, pues, serían una bisagra entre las tareas y las competencias (*vid*. figura 6), cuya taxonomía queda resumida en la tabla 9:

Competencias generales	Competencias específicas
• *Savoir* • *Savoir-faire* • *Savoir-être* • *Savoir-apprendre*	• Lingüística • Sociolingüística • Pragmática

Tabla 9. Taxonomía de las competencias propuesta por el MCER

De entre todas ellas, tal vez la más relevante para nuestro objeto de estudio sea la competencia lingüística, que, a su vez, está dividida en distintas subcompetencias: léxica, gramatical, semántica, fonológica, ortográfica y ortoépica.

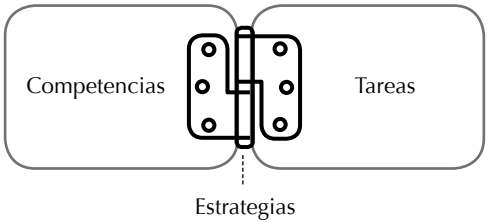


Figura 6. Relación entre competencias, estrategias y tareas

Tras la publicación del MCER (Consejo de Europa, 2001), el Consejo de Europa ha seguido apostando por la visión en él contenida, ha generado una importante cantidad de recursos teóricos y prácticos, ha fomentado desde Estrasburgo el debate en torno a la publicación original, e incluso publicó una revisión del mismo en 2020, el conocido como MCERVC (Consejo de Europa, 2020). El MCERVC (*ibid*.) es una actualización del texto de 2001 en la que se amplían muchas de las tablas originales y se incluyen otras para reflejar la evolución de la sociedad (Interacción *Online*, Lenguaje de Signos), se refuerzan determinadas ideas para las que en su momento no se facilitaron descriptores (especialmente las relacionadas con las competencias plurilingües y pluriculturales, y con la mediación) (*vid*. North 2016; North y Piccardo, 2022), se eliminan las referencias a los hablantes nativos como ejemplo de dominio, y todo ello desde una perspectiva neutral en términos de género. El MCERVC (Consejo de Europa, 2020) es la actualización necesaria, más accesible para el lector no experto, de un documento que seguirá vigente durante décadas y cuyos objetivos continúan siendo los de la publicación inicial: dotar de un lenguaje común y vertebrador a una Europa diversa, en la que se considera al hablante como un agente social que co-construye el significado a través de la interacción y la mediación en un contexto plurilingüe y pluricultural (Consejo de Europa, 2020:21).

Sin duda alguna, el impacto del MCER (Consejo de Europa, 2001) ha superado cualquier expectativa inicial de sus promotores y autores. Como decíamos, no solo se ha traducido a más de cuarenta idiomas, sino que se ha intentado replicar en países como Japón (Tono, 2019) y es usado en mayor o menor medida en múltiples contextos, como por ejemplo el hispanoamericano, donde tal vez consiga arraigar de forma definitiva como la herramienta útil que es si logra desligarse de cualquier vinculación con el colonialismo.

3.3. El argumento de validez

La validez de una prueba de dominio no se puede demostrar con un único número. No existe un test, como los que detectan el colesterol o la diabetes, que nos diga si nuestro examen es válido o no. A la validez de una prueba se llega

reuniendo evidencias sobre el rigor y la operatividad de las distintas facetas que la componen. Solo cuando reunamos todas las evidencias necesarias podremos construir el argumento que sustente la validez de nuestras pruebas.

El de la validez es un concepto esquivo cuyo carácter mutable y complejo ha variado a lo largo del tiempo (Messick, 1989; Kane, 2001; Prieto y Delgado, 2010). La noción de validez se comenzó a fraguar a mediados del siglo XIX y atravesó distintas fases (Newton y Shaw, 2014:27-61) hasta cristalizar a mediados del siglo XX en las recomendaciones técnicas de la American Psychological Association (APA, 1952) y en la que es, probablemente, la definición más conocida del término, la de Cronbach y Meehl (1955), para quienes existen tres tipos de validez: 1) la validez orientada al criterio (que a su vez puede ser predictiva o concurrente), 2) la validez de contenido y 3) la validez de constructo.

A partir del trabajo de Cronbach y Meehl (1955), la noción se fragmenta para generar una miríada de visiones distintas (validez interna, validez externa, validez de apariencia, validez cognitiva, etc.), que serán reunificadas por Messick a mediados de la década de 1970 (Newton y Shaw, 2014). Para Messick (1989:13), la validez: 1) se construye mediante distintas fuentes de evidencia científica, 2) no busca validar un examen sino las consecuencias que tienen las decisiones tomadas en función de los resultados obtenidos en él, 3) se puede dar en diferentes grados (no es cuestión de todo o nada), y 4) evoluciona con el tiempo.

Con posterioridad a Messick, entrando ya el siglo XXI, la noción de validez de nuevo se deconstruye (Newton y Shaw, 2014:135) y da lugar a modelos como el sociocognitivo de Weir (2005), el de O'Sullivan (Chalhoub-Deville y O'Sullivan, 2020) o el de Kane (2006). Si bien somos deudores de todo este debate sobre la validez, la conversación se ha enrevesado tanto en las últimas décadas que es difícil encontrar una definición operativa del término. Sea cual sea la aproximación teórica por la que nos decantemos, como Kane (1992) propuso por primera vez, las evidencias que reunamos sobre la validez de nuestras pruebas no tendrán el efecto deseado si no se presentan como un argumento coherente y convincente (O'Sullivan, 2014:27).

Dada esta variedad de perspectivas, el argumento de validez de una prueba puede tener formas muy diversas. Para ALTE (Association of Language Testers in Europe), por ejemplo, un argumento de validez ha de contener 1) vinculaciones explícitas entre el sistema que se desea validar y los diecisiete estándares mínimos de calidad establecidos por la propia asociación (relacionados con la construcción del examen, la administración y logística, la forma de calificar las pruebas, los sistemas de análisis utilizados y la comunicación con los agentes interesados). Además, 2) el argumento de validez ha de aportar información sobre estas vinculaciones, 3) y ha de hacerlo de manera debidamente justificada mediante 4) las pruebas necesarias (especificaciones, análisis psicométricos, informes técnicos sobre requisitos de instalaciones, *software* o *hardware*), etc. (ALTE, 2020:25).

Una pregunta lógica sería, pues, ¿qué forma debe tener el argumento de validez de mis pruebas? La respuesta a esta pregunta depende del contexto en el que se plantee. Con frecuencia, las entidades confeccionadoras de exámenes han de armonizar aspectos tan distantes entre sí como la lingüística, la psicometría y las presiones institucionales, económicas, sociales y políticas (Spolsky, 1995:4). Para intentar equilibrar los diferentes lados de esta compleja balanza, en la tabla 10 proponemos un esquema de argumento de validez basado en Messick (1989) y Wolfe y Smith (2007a, 2007b) que, en función del contexto y de los recursos disponibles, puede adoptarse total o parcialmente. Esta es, por tanto, una tabla orientativa, no prescriptiva, y ha de ser considerada como una guía, no como un objetivo en sí misma.

Aspecto	¿Qué evidencia este aspecto?	¿Qué pruebas lo sustentan?
Contenido	Si el contenido en el que se basa una prueba es representativo y relevante y si los ítems que la componen son de calidad	• Documentación del propósito de nuestras prueba • Documentación sobre cómo se diseñó el constructo • Documentación sobre cómo se diseñó la prueba • Revisión de expertos de los puntos anteriores • Evidencia sobre la calidad técnica de los ítems (correlación ítem-test; valores de ajuste al modelo)
Sustantivo	Si nuestros candidatos responden a los ítems con las respuestas y los procesos cognitivos que esperábamos según el constructo con el que la prueba fue diseñada	• Análisis de los procesos cognitivos de los candidatos • Observación de los patrones conductuales de los candidatos • Análisis de distractores en preguntas dicotómicas • Ajuste de los patrones de respuestas de los candidatos *(person fit)* • Confirmación del orden teórico de los ítems (mapa de la variable) • Análisis del funcionamiento de las escalas analíticas

Estructural	Si la relación estructural entre los diferentes componentes de nuestra prueba (ítems, tareas, destrezas) refleja la forma en la que la lengua se estructura	• Análisis factorial • Comprobación de la unidimensionalidad
Generalización	Si los resultados obtenidos medinte una prueba son estables en diferentes contextos	• Evidencias de invarianza de los ítems (DIF y valores de ajuste al modelo) • Evidencias de inependencia contextual (anclaje horizontal de pruebas, correlación calificador-resto de calificadores) • Fiabilidad (coeficiente de fiabilidad, *inter-rater reliability*, *intra-rater reliability*, *person separation reliability*) • Documentación derivada de sesiones de estandarización de correctores
Externo	Si nuestras mediciones se pueden relacionar con mediciones externas del mismo constructo, de constructos similares o de otros constructos; si nuestro examen es sensible a los cambios en el nivel de los canidatos; si la distribución de los ítems cubre un rango suficiente del rasgo medido; cuántos grupos de candidatos con niveles significativamente distintos diferencia nuestra prueba	• Matrices multirasgo-multimétodo • Distribución adecuada de ítems (mapa de la variable, número de estratos

Impacto	Si existe información suficiente para que las calificaciones de nuestras pruebas se interpreten de forma correcta por parte de terceros; si las consecuencias derivadas de dichas interpretaciones son las adecuadas y si existen protocolos destinados a corregir la mala praxis	• Documentación que evidencie que los usuarios de nuestras pruebas reciben información adecuada sobre el contexto, el propósito, el contenido y la fiabilidad de nuestras pruebas (disponibilidad de las especificaciones, disponibilidad de versiones inteligibles del argumento de validez, folletos, manuales, *webs* accesibles y claras, organización de seminarios, encuestas de percepción, etc.) • Documentación sobre cómo se informa a los usuarios acerca de la correcta interpretación de los resultados (informes psicométricos orientados a los responsables de la toma de decisiones, campañas de difusión, manuales, organización de seminarios, encuestas de percepción, etc.) • Documentación sobre cómo se actúa en caso de que las calificaciones de nuestras pruebas se estén usando con un propósito que va en contra del interés de los candidatos (listados de acciones consideradas como mala praxis, acciones compensatorias, etc.)

Tabla 10. Aspectos y evidencias de un argumento de validez

La tabla 10 indica qué aspectos de nuestras pruebas se deben analizar, las evidencias que el análisis de estos aspectos nos aporta y, finalmente, cómo se puede llegar a dichas evidencias.

Según este modelo, lo primero que deberíamos analizar es la relevancia y la representatividad del contenido en el que se basan nuestras pruebas (Wolfe y Smith 2007b:205–207). Para esto es particularmente interesante analizar la calidad técnica de los ítems, por ejemplo, mediante el análisis de los índices de correlación ítem-test (que Wolfe y Smith (*ibid.*) denominan correlación *item-measure*) y mediante el análisis de los valores de ajuste al modelo, de los que hemos hablado en las secciones 2.1.1.3 y 2.1.2.2 respectivamente.

En segundo lugar, el componente sustantivo de un argumento de validez es el que relaciona el planteamiento teórico de nuestras pruebas con las respuestas reales obtenidas en el examen (Wolfe y Smith 2007b: 207–213). Es decir, ¿están nuestros candidatos respondiendo a los ítems con las respuestas y los procesos cognitivos que esperábamos según nuestro constructo de lengua? Para realizar estas comprobaciones se pueden usar herramientas como el análisis de los procesos cognitivos de los candidatos (mediante entrevistas en las que se les pregunte cómo afrontan la prueba); la observación de patrones conductuales (por ejemplo, analizando cómo leen los exámenes (Brunfaut, 2022)); el análisis de distractores en preguntas dicotómicas (Green, 2013:185-193); el ajuste de los patrones de respuestas de los candidatos (o *person fit*, que serían los valores de ajuste descritos en 2.1.2.2 aplicados a los candidatos); la confirmación del orden teórico de los ítems (reflejada en el mapa de la variable descrito en la sección 2.1.2.3); y el análisis del funcionamiento de las escalas analíticas (a las que dedicaremos el capítulo 4 por completo).

El siguiente aspecto que debería analizar nuestro argumento de validez es el estructural (Loevinger 1957:661; Wolfe y Smith 2007a, 2007b). Para analizar la vinculación entre la estructura de los componentes de nuestras pruebas (observados a través de nuestro sistema de puntuación) y la estructura de otras manifestaciones del atributo medido (por ejemplo, la comprensión lectora), se puede realizar un análisis factorial y se debe contrastar la unidimensionalidad. Por lo general, nuestros exámenes de dominio de lengua están diseñados para medir un único atributo, una única dimensión (comprensión lectora, comprensión auditiva, etc.). Los resultados de nuestras pruebas están, no obstante, compuestos por dicha dimensión y por elementos que, no perteneciendo a ella, pueden distorsionar la interpretación de los datos. Estos elementos distorsionadores o «residuos», están compuestos por ruido aleatorio y por trazas de interconexión entre ítems. Si dichos indicios de interconexión son excesivos podrían indicar que nuestros ítems están midiendo una segunda dimensión (por ejemplo, podríamos estar midiendo algo diferente a la comprensión auditiva si penalizamos las faltas de ortografía en determinados niveles de dominio). Para saber si existe una o más dimensiones en nuestros datos se usa el método de análisis de componentes principales, que tiene en cuenta el número de candidatos e ítems (Chou y Wang, 2010). Conviene recordar que siempre encontraremos trazas de dimensiones secundarias en nuestros exámenes, dado que la complejidad de la existencia humana no puede ser expresada de forma satisfactoria mediante la calificación obtenida en ningún examen, por bien construido que este esté (Bond y Fox 2007:33–34). Por lo tanto, al analizar la integridad estructural de nuestras pruebas la verdadera pregunta que hemos de hacernos es dónde está el límite que justifique (o no) la creación de exámenes distintos para cada una de las posibles subdimensiones que encontremos en ellas (*vid*. Linacre, 2024a).

La capacidad generalizadora de nuestro argumento de validez analiza el grado en el que las mediciones que realizamos se mantienen estables en diferentes contextos, es decir, si los valores que obtenemos de nuestros ítems y tareas son los mismos cuando la prueba la realizan distintos candidatos, cuando se realiza en distintos momentos (Wolfe y Smith, 2007b:215) o en distintos formatos (papel frente a soporte digital). Esto está estrechamente ligado con la invarianza que ya quedó descrita en la sección 2.1.2 y con el análisis del funcionamiento diferencial de ítems (o DIF, del inglés *differential item functioning*), que nos indica si hay sesgo de algún tipo en nuestras pruebas (Gómez-Benito *et al.*, 2010). Mediante el análisis del sesgo, por ejemplo, podremos saber si dada la misma habilidad en la variable latente la probabilidad de acierto en un ítem es mayor para distintos grupos de candidatos (hombres frente a mujeres; hablantes de lenguas romances frente a los de otras lenguas en un examen de español, etc.). La capacidad generalizadora de nuestros exámenes también quedará demostrada si sus distintas versiones tienen el mismo nivel de dificultad, algo que se puede analizar mediante el anclaje horizontal de pruebas (Wolfe, 2004) y que nos servirá para contrastar si, por ejemplo, la convocatoria de exámenes de nivel B1 diseñada un año tiene la misma dificultad que la diseñada al año siguiente. También podremos considerar la fiabilidad y discriminación de los ítems (*vid.* 2.1.1.1 y 2.1.1.3), la consistencia de las calificaciones de los evaluadores (de la que hablaremos en el apartado i de la sección 4.2.4.2) o la fiabilidad de la ordenación de los candidatos (Bond y Fox, 2007:40–41).

El aspecto externo de nuestro argumento de validez es, según Wolfe y Smith (2007b:220), uno de los más importantes y el que más cercano está a la idea tradicional de validez de constructo de Cronbach y Meehl (1955). Lo que investiga este aspecto es hasta qué punto las mediciones obtenidas mediante nuestros exámenes pueden relacionarse con otras medidas (externas) del mismo constructo, de constructos similares o de constructos diferentes (por ejemplo, ¿tienen los exámenes de nivel C1 de dos escuelas oficiales de idiomas de comunidades distintas el mismo nivel de dificultad?). Para investigar el aspecto externo de nuestras pruebas se pueden usar matrices multirasgo-multimétodo (Campbell y Fiske, 1959), se puede analizar el mapa de la variable (*vid.* 2.1.2.3) e investigar a cuántos estratos de habilidad distintos es sensible un examen (Wright y Masters, 1982:106), siendo dos (aprobado frente a suspenso) el mínimo deseable en un examen uninivel.

Por último, nuestro argumento de validez ha de reflexionar sobre la repercusión que un examen tiene en su contexto. Con frecuencia, las calificaciones de nuestros candidatos trascienden a la prueba en sí y son utilizadas por terceros en la toma de decisiones. Pensemos en los exámenes de acceso a la universidad, en los exámenes para obtener la nacionalidad o en la ventaja que puede suponer en unas oposiciones poseer el certificado de una escuela oficial de idiomas. Desde esta perspectiva, la responsabilidad del impacto de una prueba descansa funda-

mentalmente en la entidad confeccionadora, que es la encargada de transmitir la información necesaria y de velar por que esta sea interpretada de forma correcta. No se ha de caer, pues, en el error de pensar que nuestra responsabilidad institucional acaba una vez que se hacen públicas las calificaciones.

3.4. Componentes principales

Llegamos ahora a un punto delicado en el desarrollo de esta publicación. Para dar forma a cualquier examen de dominio es necesario descomponer la gran complejidad del lenguaje humano en elementos separados. Como vimos en el capítulo 1, esto nos conduce a debates de calado sobre la naturaleza del lenguaje, el objeto que deseamos observar. Es lícito cuestionar si tiene sentido intentar medir el lenguaje cuando no existe un consenso universal sobre lo que este es. Tan lícito y científico es plantearse estas dudas como lo es lanzarse en busca de respuestas, aunque sean parciales, y aunque estas procedan de manifestaciones indirectas del lenguaje.

Por cuestiones prácticas, en las secciones que siguen describiremos los que tradicionalmente han sido considerados como componentes principales del lenguaje (comprensión auditiva, comprensión lectora, producción e interacción escritas y producción e interacción orales). Somos conscientes de que este modelo, originado en la segunda mitad del siglo XX (como se nos recuerda en el prólogo de este libro), actualmente está siendo revisado. Una de estas revisiones, particularmente relevante dentro del contexto europeo, ha sido precisamente la que ha venido proponiendo el MCER (Consejo de Europa, 2001) desde 2001 y que se ha ampliado con el MCERVC (Consejo de Europa, 2020).

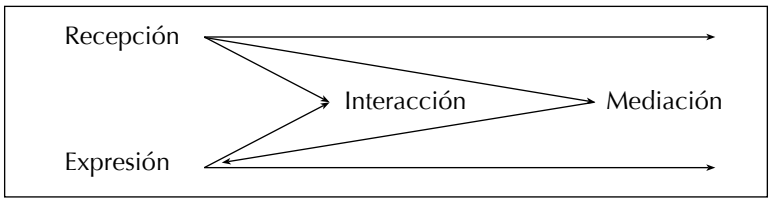


Figura 7. Modos de comunicación según el MCER y el MCERVC

El modelo del MCER (Consejo de Europa, 2001) y del MCERVC (Consejo de Europa, 2020) transita desde las tradicionales destrezas productivas y receptivas a una visión más amplia del lenguaje que trasciende incluso la integración de dichas destrezas. El posible diseño de exámenes basados en este modelo obliga no solo a reflexionar sobre los componentes de nuestra prueba, sino también sobre la herramienta mediante las que estos se articulan: las tareas.

Tradicionalmente se ha intentado crear tareas funcionales para la evaluación que reducían mucho el constructo, lo que con frecuencia ha hecho que aquellas se separen de este. Un cambio de paradigma en los componentes de nuestras pruebas tal vez pueda ayudarnos a avanzar para encontrar una solución de compromiso entre la artificialidad de cualquier examen y un modelo de tarea que capture mejor la realidad de la comunicación lingüística al tiempo que mantiene su funcionalidad. Todo ello puede suponer profundos cambios en las pruebas, como se ha visto en Grecia (Stathopoulou, 2018), Austria (Atzlesberger, 2015) o España (Cruz, 2024), donde se ha experimentado con la integración de la mediación. Estos nuevos modelos son tan interesantes como complejos, dado que plantean interrogantes como el de la operatividad de las pruebas: ¿Es factible o costeable evaluar la mediación interlingüística en un examen de dominio en el que participen dos o más interlocutores? ¿Cómo se diseña una tarea para este objetivo, adecuadamente contextualizada, que indique de forma clara a los candidatos lo que se espera de ellos y cuya clave esté definida de forma fiable? ¿Cómo se miden las contribuciones de los distintos candidatos? ¿Se han de evaluar exclusivamente las estrategias que usen los candidatos, el producto final o ambas cosas? ¿Se puede/debe utilizar más de una lengua en este tipo de exámenes de dominio? En tal caso, ¿cuántas lenguas se han de utilizar? ¿Es la L1 la lengua nativa del candidato y la L2 la segunda lengua? ¿Qué ocurre cuando L1 y L2 no coinciden en todos los candidatos? ¿Tiene sentido incluir una L3?, etc.

En las secciones que siguen iremos de lo teórico y general a lo concreto y práctico. Siguiendo una taxonomía tradicional, primero hablaremos del constructo y las especificaciones de nuestros exámenes, para luego aterrizar en los distintos componentes que constituyen un examen (comprensión auditiva, comprensión lectora, producción e interacción escritas y producción e interacción orales) que se articulan mediante tareas. Aunque no perdemos de vista modelos distintos al tradicional, y siendo conscientes de los interesantes cambios que los avances tecnológicos operarán en la evaluación de lenguas, entendemos que dichos modelos no han alcanzado aún la madurez necesaria ni han generado el consenso requerido para desplazar al paradigma tradicional de cuatro destrezas.

3.4.1. Constructo y especificaciones

El término constructo se usa con frecuencia como una especie de cajón de sastre que agrupa todas las características de nuestras pruebas. Este uso, intencionadamente ambiguo, si bien es susceptible de reflejar constantemente nuevos matices, es poco útil. El constructo es a nuestros exámenes lo que el nivel del mar es a la geografía, es decir, la línea de referencia desde la que se empieza a dibujar (*vid*. Ortega y Gasset, 1956:52), por lo que, usado de forma ambigua, es una herramienta inútil.

De formas similares, Bachman (1991:41–44), Fulcher y Davidson (2007:7) y Kerlinger y Lee (2000:40) nos sugieren que los constructos han de tener dos características fundamentales: estar definidos de forma teórica y ser operativos. Al definir un constructo teóricamente podemos relacionarlo con otros y al hacerlo operativo lo podemos observar y medir.

Definir teóricamente un constructo significa caracterizar aquello que se desea medir, en nuestro caso, el lenguaje. El lenguaje, al contrario que otras características físicas como la altura o el peso, no puede ser observada de forma directa, como tampoco pueden ser observados el amor o el pensamiento. Por lo tanto, al definir nuestro constructo hemos de establecer de forma clara cuáles son sus características, algo que como ya hemos visto en el capítulo 1, hoy por hoy solo puede hacerse de forma parcial. Por ejemplo, un constructo de lenguaje basado en el MCER (Consejo de Europa, 2001) y en el MCERVC (Consejo de Europa, 2020) partiría de los modos de comunicación de la figura 7 de más arriba, en la que el hablante es concebido como un agente social que emplea determinadas estrategias para activar competencias en un uso de la lengua orientado a la acción y al desarrollo de tareas que se dan en distintos dominios.

Hacer nuestro constructo operativo, por otro lado, supone encontrar la manera de hacer sus características observables y susceptibles de ser comparadas con otras mediciones. Para hacer operativo el constructo que hemos definido en el párrafo anterior, por ejemplo, podríamos diseñar una tarea de producción e interacción orales (modos de comunicación) en la que dos candidatos asumen el rol de estudiantes universitarios (dominio académico) y debaten entre sí acerca de posibles mejoras arquitectónicas para el campus en el que estudian (agente social). La producción de estos candidatos podría medirse con unas escalas diseñadas para identificar qué estrategias emplean para activar distintas competencias (lingüísticas, sociolingüísticas y pragmáticas).

Generalmente, ninguna prueba está vinculada a un documento que lleve por título «Constructo». El constructo de una prueba se puede encontrar diseminado en los distintos documentos que utilizan los redactores de pruebas, los correctores y los candidatos. Entre estos documentos están las especificaciones de las pruebas, que deben comenzar con la definición teórica del constructo (que tendrá diferente nivel de detalle según la audiencia a la que esté dirigida) y luego indicar cómo este se hace este operativo. Por ejemplo, en el caso de unas especificaciones para redactores de pruebas, se ha de establecer claramente qué componentes tiene el examen, el tiempo que los candidatos tendrán que emplear en cada uno, el número de tareas e ítems, las estrategias y competencias que cada tarea busca activar y la forma en que se presentarán dichas tareas (nivel de las instrucciones, contextualización y dominio de la tarea, ejemplos, tipo de textos, vocabulario tabú, etc.). Este tipo de especificaciones ha de ser una guía clara que permita diseñar tareas que generen en los candidatos un uso

de la lengua estandarizado y en condiciones uniformes (Bachman, 1991:43). Las especificaciones destinadas exclusivamente a los redactores de pruebas suelen ser de carácter interno.

Las especificaciones que sí pertenecen siempre al dominio público y que también contienen el constructo de la prueba son aquellas orientadas a los candidatos. Suelen tener nombres como *guía del usuario*, *manual del candidato*, *guías de examen* o similares, y contienen descripciones exactas del número de tareas e ítems en cada componente, indican la duración de la prueba, y suelen estar acompañadas de modelos de examen resueltos, de hojas de respuesta, de las escalas de corrección utilizadas por los evaluadores, etc. Ejemplos de ello son las guías de examen DELE del Instituto Cervantes (2014a; 2014b; 2019; 2022) o las de la *suite* de exámenes Cambridge University Press and Assessment (2024a; 2024b; 2024c; 2024d). Estos documentos están orientados a dotar de transparencia a las pruebas y a ser un «manual de instrucciones» claro para los candidatos, por lo que dedican más atención a describir cómo se hace operativo el constructo que a definirlo teóricamente.

Por último, cabe mencionar que cualquier constructo es el trasunto de una necesidad que, en última instancia, es la que demanda la construcción de una prueba concreta con características específicas. Como veremos más adelante, no es lo mismo diseñar una prueba de dominio para controladores aéreos que diseñar una prueba para estudiantes universitarios.

3.4.2. Comprensión auditiva

Al igual que ocurre con todas las destrezas que analizaremos, la comprensión auditiva cuenta con sus propias características (*vid*. Green, 2017:1–19 para un análisis detallado de dichas características).

El oído es la principal puerta de entrada de *input* lingüístico al cerebro. Es imposible aprender a hablar si antes no se aprende a escuchar. Cuando escuchamos, las palabras nos llegan en un flujo continuo en el que no existen espacios en blanco. Con frecuencia, los hablantes noveles de una lengua refieren la dificultad que les supone esta carencia de barreras entre palabras habladas, algo que hemos experimentado todos los que aprendimos una segunda lengua en la edad adulta. Este flujo constante, además, impide que podamos recorrer en sentido inverso el mensaje acústico más allá de lo que nos permita retener nuestra memoria de trabajo. Por todo ello, la comprensión auditiva es, junto con la producción oral, uno de los componentes que más estrés suelen generar en nuestros candidatos.

Son múltiples los factores que se han de tener en cuenta al diseñar pruebas de comprensión auditiva. A continuación, partiendo de Green (2017), exponemos una serie de recomendaciones sobre:

- El lugar de celebración de la prueba
  - Características acústicas de la sala
  - Disposición de los candidatos
  - Características de los equipos de reproducción

- El archivo de audio
  - Autenticidad y guionización
  - Contenido
  - Calidad del sonido
  - Número y longitud de las grabaciones
  - Número y tipo de voces
  - Velocidad de dicción
  - Uso de imágenes

- La tarea
  - Instrucciones y ejemplo
  - Tipo de examen
  - Número de ítems y de tareas

Comenzando por lo más obvio, el lugar donde se haya de celebrar la prueba ha de ser tenido en cuenta y revisado con una antelación tal que nos permita hacer cambios (de equipo, de ubicación, etc.) y comunicárselos a los candidatos en caso de ser necesario. La capacidad de un hablante de poner en acción su competencia comunicativa depende en gran medida de las condiciones físicas en las que esta tiene lugar (Consejo de Europa, 2001:46–47). Así pues, las características acústicas de la sala en que se realiza la prueba son de vital importancia y han de ser comprobadas al menos en tres ocasiones: días antes de la prueba, momentos antes del comienzo de la prueba con la sala vacía y justo al comienzo de la prueba una vez que los candidatos ya están dentro.

Durante las comprobaciones en días anteriores a la prueba es recomendable reproducir el audio del examen en la sala sin candidatos y desplazarse por ella para comprobar que la calidad es uniforme y que el lugar donde un candidato se siente no afectará a su rendimiento en la prueba. Si un candidato está ubicado demasiado cerca de una de las fuentes de sonido, esta puede saturar o molestar durante la escucha. Si está ubicado demasiado lejos, la señal acústica puede ser débil. Para evitar esto es recomendable contar con varias fuentes de sonido ecualizables (es decir, en las que se pueda modificar volumen, balance, graves, agudos, etc.). También se ha de comprobar si la sala produce ecos, reverberaciones que hagan que el sonido se solape o se acople. El uso de auriculares conectados a una misma unidad de distribución puede garantizar una calidad de audio uniforme,

como ocurre por ejemplo en el examen BCT-S, una versión del examen *Aptis* del British Council diseñada conjuntamente con la Tokyo University of Foreign Studies. Si se opta por este sistema, se habrá de comprobar que los auriculares funcionan de forma correcta para evitar problemas de reproducción justo antes del comienzo de la prueba (conexiones mediante cable, *Bluetooth*, etc.). Las comprobaciones habrán de realizarse exactamente con el mismo soporte que se utilizará el día de la prueba. Así, por ejemplo, si la pista de audio ha de descargarse de forma remota en un ordenador, esta ha de descargarse de la misma manera durante la comprobación. Si el audio se va a reproducir desde una memoria externa, se ha de reproducir desde dicha memoria para comprobar que esta funciona, etc.

Además de la comprobación de sonido realizada en días previos, es recomendable llevar a cabo una segunda comprobación el día del examen antes de que los candidatos accedan a la sala. Realizaremos esta comprobación para cerciorarnos de que en los días transcurridos desde la primera no ha habido ningún cambio en los equipos. En ocasiones las salas de las pruebas se usan para diversos fines, y es posible que se haya modificado la ecualización de los equipos, la distribución de las fuentes de sonido, etc. Esta segunda comprobación requerirá menos tiempo que la primera, ya que las modificaciones de calado debieron quedar realizadas durante la primera comprobación.

La tercera de las comprobaciones está destinada a los candidatos. Como ya hemos comentado, la comprensión auditiva genera incertidumbre y ansiedad entre los candidatos. La experiencia nos dice que durante los primeros segundos de cualquier grabación los candidatos están aún acomodándose al tono, al volumen, al ritmo de la grabación, y por ello es recomendable que, con los candidatos ya ubicados en la sala, después de pedirles que desconecten cualquier tipo de dispositivo que pueda producir ruido o interferencias durante la prueba, reproduzcamos los primeros quince o veinte segundos de instrucciones de la grabación anteriores al comienzo de la primera tarea. Esto ayudará a determinar si la acústica de la sala cambia cuando está llena, al tiempo que dará a los candidatos la oportunidad de sugerir pequeñas modificaciones en el volumen o la ecualización. Generalmente, tener este control del sonido predispone de forma positiva a los candidatos. Transcurridos estos quince o veinte segundos y hechas las modificaciones oportunas (si son necesarias), se podrá proceder a reproducir la pista de audio desde el principio dando así comienzo a la prueba de comprensión auditiva. El mismo tipo de comprobación se puede aplicar a las pruebas realizadas mediante distintos auriculares conectados a una misma fuente de distribución.

El segundo conjunto de factores importantes en el desarrollo de pruebas de comprensión auditiva es el relacionado con los archivos de audio que utilizaremos. Estos audios deberán ser, en la medida de lo posible, auténticos (*vid.* 3.5.1), es decir, deberán contener textos orales que puedan darse en el contexto en el

que se enmarca la prueba (*vid*. Lewkowicz, 1996; 2000 y Green, 2017:37–38). Así, por ejemplo, si estamos diseñando tareas para un examen destinado a controladores aéreos, debemos proponer audios que puedan darse en una situación cotidiana en la que un controlador deba, pongamos por caso, asignar un *slot* de aterrizaje a una aeronave, indicar las condiciones cambiantes del viento, etc., mientras que si estamos diseñando un examen de dominio para un contexto universitario deberíamos proponer tareas en las que, por ejemplo, un profesor se dirija a sus alumnos durante una lección, o en la que se escuche a un estudiante solicitando información sobre el sistema de préstamo de una biblioteca, etc. Encontrar audios realistas, algo que será más difícil en niveles de dominio inferiores, reforzará la validez de apariencia de nuestras pruebas. Conseguir esta autenticidad no siempre es sencillo. Con frecuencia, la autenticidad está reñida con las características que hacen de un audio una buena opción para determinados exámenes. Siguiendo el último ejemplo, imaginemos que grabamos una conversación auténtica entre una estudiante y el bibliotecario que le presta servicio. En esta conversación espontánea los hablantes no adaptarán su velocidad de dicción a un nivel determinado, no limitarán su vocabulario a un campo específico, no distribuirán de forma heterogénea ideas principales y secundarias a lo largo de su conversación ni se adaptarán a una duración determinada. La conversación es auténtica, pero puede no ser adecuada para nuestros fines si la usamos tal cual es registrada. Será importante decidir si para la preparación de los audios nos limitaremos a grabar textos surgidos de forma más o menos natural en una conversación o si, por el contrario, prepararemos de antemano un guion para que un locutor realice una grabación que se adapte a las especificaciones de nuestra prueba.

Tanto las conversaciones espontáneas como las guionizadas pueden ser válidas para nuestras pruebas si mantienen la autenticidad de la que antes hablábamos. Con la grabación de un audio previamente guionizado no solo podemos conseguir que los intervalos de tiempo entre ideas sean regulares, sino que también podremos incluir un número determinado de distractores que nos ayuden a generar ítems más equilibrados. Cambridge University Press and Assessment, por ejemplo, utiliza grabaciones guionizadas para mantener una calidad constante en sus pruebas de comprensión auditiva. La guionización que hagamos de la grabación podrá afectar a mayor o menor parte de su contenido. Si se realiza una guionización completa se corre el riesgo de convertir el texto auditivo en la mera lectura de un texto escrito, algo que, si bien puede ocurrir en la vida real (pensemos en el discurso de un político, en las llamadas y mensajes que se escuchan en los aeropuertos o en las previsiones del tiempo que se emiten en los boletines radiofónicos), es menos frecuente que una conversación espontánea. Para evitar, si se desea, esta guionización completa se puede, por ejemplo, proponer un tema concreto al locutor o locutores y una serie de hitos, temas o palabras a las que

han de hacer referencia. Por ejemplo, a la hora de realizar la grabación de un diálogo sobre las vacaciones de dos personas se les podría pedir que durante la conversación hicieran ambos alusión a cuáles han sido las mejores vacaciones de su vida, cuáles han sido las últimas y cuáles las que aún no han tenido pero les gustaría tener. También se les podría suministrar una lista de palabras concretas que se han de mencionar y que más tarde pueden servir para la redacción de determinados distractores. De igual manera, se les podría pedir que dedicasen un tiempo concreto a tratar cada una de las cuestiones para que durante la redacción de ítems sea más sencillo establecer intervalos regulares entre ellos. Si se ejecuta de forma correcta, una grabación semiguionizada puede dar lugar a audios con un buen equilibrio entre espontaneidad y recursos para el diseño de ítems. Ninguno de estos tres tipos de registros sonoros (espontáneos, guionizados o semiguionizados) es intrínsecamente superior al resto y todos son igualmente válidos si, como decíamos, preservan su autenticidad.

El contenido del texto sonoro de nuestras tareas estará determinado por el constructo de la prueba. Como ya hemos indicado, los registros sonoros habrán de estar relacionados con aquellas tareas en las que los candidatos tendrían que hacer valer su dominio de la lengua en una situación real. Algunos exámenes crean varias versiones de una misma prueba para reflejar diferentes dominios. IELTS, por ejemplo, tiene una versión general y otra académica. Algunas de las pruebas de la *suite* de exámenes de Cambridge University Press and Assessment tienen una versión para escolares y otra para adultos. Hemos de buscar archivos de audio que no contengan referencias culturales específicas o locales, que puedan beneficiar a unos pocos candidatos en detrimento del resto, al tiempo que hemos de contextualizar en la medida de lo posible la temática del audio en las instrucciones de la tarea (Green, 2017:14–15).

Con anterioridad hemos mencionado la importancia de la acústica de la sala donde se realiza el examen y de cómo los equipos de reproducción o la disposición de los candidatos pueden afectar el desempeño de estos últimos. Previo a todo ello hemos de tener en cuenta que la calidad del sonido de las grabaciones habrá de ser óptima. Por muy bien equipada que esté nuestra sala de examen, si el audio que reproducimos está dañado, satura o distorsiona, los candidatos sufrirán las consecuencias de una mala planificación. Es importante realizar las grabaciones sin sonidos de fondo que interfieran en la comprensión, salvo que dicho sonido de fondo esté considerado en nuestro constructo y especificaciones (por ejemplo, el CEFR establece que el usuario B2 de la lengua ha de ser capaz de entender lo que se dice a pesar del ruido de fondo (Consejo de Europa, 2001:66; 234)).

Al seleccionar archivos de audio procedentes de fuentes externas como internet, nos hemos de cerciorar de que su calidad cumple unos requisitos mínimos. De la misma manera, cuando los archivos de audio son de producción propia se

ha de intentar grabarlos en un estudio profesional en el que se pueda controlar tanto el ruido como la ecualización de los diferentes interlocutores. Si esto no es posible, se puede considerar adquirir un equipo de grabación semiprofesional que, en un ambiente controlado, puede dar resultados igualmente buenos. Muchos dispositivos digitales ofrecen hoy una calidad de grabación equiparable a la de los mejores equipos de hace años. Una vez seleccionadas o realizadas las grabaciones, existen multitud de programas gratuitos que permiten la edición de audio (recorte de tomas falsas, ecualización, modulación de ganancia, introducción de efectos para separar o contextualizar tareas, etc.) de forma muy sencilla. Nosotros recomendamos utilizar *Audacity* (Audacity Team, 2024), un *software* que ha adquirido popularidad en los últimos años por ser gratuito y muy sencillo de utilizar. No obstante, existen otros programas de este tipo igualmente válidos y, a buen seguro, en el futuro se desarrollarán aún más. La experiencia nos dice que es recomendable sopesar detenidamente qué *software* se va a usar para que, una vez tomada la decisión, esta pueda mantenerse durante años y evitar así problemas de compatibilidad entre archivos al pasar de un sistema a otro, una recomendación que hacemos extensiva a todos los programas de edición que se usen en el diseño y maquetación de las pruebas.

El número y longitud de las grabaciones, una vez más, estarán determinados por nuestro constructo que, a su vez, estará articulado en nuestras especificaciones. Como hemos indicado en la sección 3.4.1, las especificaciones han de ser accesibles para los candidatos de forma que estos puedan conocer con exactitud antes de la prueba cuál es la estructura de la misma. Es de suma importancia que las especificaciones establezcan la duración mínima y máxima de los audios para que tanto los redactores de las pruebas como los candidatos a las mismas sepan a qué atenerse. Distintos tipos de tarea pueden requerir distintos tipos de audio que, a su vez, podrán ser más o menos complejos y extensos en función del nivel que evaluemos. Por lo general, se usará un único archivo de audio de una duración media o extensa o bien una combinación de audios más cortos. Si se usan varios extractos, estos pueden guardar relación entre sí (e.g., las diferentes partes del discurso de un mismo orador sobre un tema concreto) o no (e.g., diferentes cortes de un boletín radiofónico que hacen referencia a distintas noticias). En este sentido, por ejemplo, Green (2017:41) recomienda usar varios archivos de audio porque que esto permite exponer a los candidatos a diferentes estructuras discursivas y porque cada archivo de audio ofrece al candidato una nueva oportunidad para demostrar su capacidad de comprensión. Utilizar varios archivos de audio también permite al redactor, si es adecuado para los propósitos de la tarea, definir distintos bloques temáticos dentro de la misma.

En ocasiones será más conveniente utilizar una sola voz y en otras será preferible usar varias. Tanto en las especificaciones para redactores como en los ma-

nuales para candidatos se ha de incluir información sobre este número de voces y sobre si se usará una sola variedad dialectal o distintas. Hemos de pensar que, en algunos contextos, los candidatos pueden haber estado expuestos a un número muy limitado de voces y acentos antes de la prueba. Si bien este hecho no ha de condicionar el diseño de nuestros exámenes, es sin duda algo que conviene tener en cuenta (Green, 2017:17).

Aunque en el aula la exposición de los alumnos a una velocidad de dicción natural puede mejorar su capacidad de comprensión auditiva, se ha demostrado que una velocidad de dicción superior a la media reduce las calificaciones de los candidatos en un examen y que una velocidad inferior a la media las aumenta (Griffiths, 1992). Estas ideas, aparentemente contradictorias, parecen indicar que lo que es bueno en el aula no tiene por qué serlo para el diseño de nuestras pruebas. Cabe pensar que la L1 del candidato también pueda afectar a su capacidad de comprender un determinado número de palabras por minuto (por ejemplo, puede que para los hablantes de lenguas romances sea más rápido el procesado de palabras en un examen de español de lo que lo sería para candidatos de lenguas germánicas o asiáticas). Todos estos son factores que, si bien no deben condicionar el diseño de nuestras pruebas, han de ser tenidos en cuenta. Sería recomendable, pues, determinar en las especificaciones cuál se considera la velocidad media de dicción y tomarla como referencia para disminuirla, mantenerla o aumentarla según sea necesario en diferentes niveles. Para un análisis más amplio puede consultarse Green (2017:17–18; 45–46).

Por último, habremos de considerar si la tarea se acompaña o no de imágenes. Aunque los recursos visuales son parte de la vida cotidiana, y por tanto pueden aportar realismo a nuestras tareas, en caso de que se decida utilizarlos se habrá de cuidar su edición y presentación. Si, por ejemplo, el color de la imagen es importante para la correcta realización de la tarea tendremos que garantizar que la versión final del examen muestre dichos colores de forma adecuada (i.e., no se presentan en blanco y negro, la pantalla no altera la percepción de los colores, etc.).

3.4.2.1. Las tareas

Continuando nuestro recorrido, abandonamos lo general y llegamos a lo particular para definir lo que es una tarea. Si nuestro constructo se articula en las especificaciones, estas últimas se articulan en las tareas. Una tarea es cualquier actividad de la vida diaria en la que utilizamos nuestras competencias (generales y específicas) mediante estrategias (receptivas, productivas, de interacción o de mediación) con el fin de lograr un objetivo dentro del ámbito (o dominio) personal, público, educativo o profesional (Consejo de Europa, 2001:157). Por ejemplo,

llevamos a cabo una tarea cuando en un viaje de trabajo (dominio profesional), leemos la pantalla de información que anuncia nuestro vuelo. En este caso somos capaces de determinar a qué puerta de embarque dirigirnos gracias a que una estrategia receptiva (la comprensión lectora) activa nuestro conocimiento sobre el funcionamiento de los aeropuertos (competencias generales) y sobre el idioma en el que los anuncios están escritos (competencias comunicativas).

Las tareas están compuestas por instrucciones e ítems, y estos últimos compuestos a su vez por una raíz, la clave y, si es necesario, distractores. Mediante las instrucciones dotamos a nuestras tareas de similitud con la vida real, contextualizándolas sucinta y debidamente con el objeto de activar determinados esquemas mentales en los candidatos. Algunos redactores consideran innecesario que en las instrucciones de una tarea se proponga una situación imaginaria que la contextualice. En nuestra opinión, sin embargo, vincular cualquier tarea con una actividad de la vida diaria ayuda al candidato a identificarse con los objetivos de la tarea y le predispone de manera positiva ante la misma. Con todo, por convincente y natural que parezca cualquier tarea, esta siempre será un artefacto diseñado para que los candidatos participen en situaciones ficticias. Por este motivo es importante incluir ejemplos en las instrucciones. Los ejemplos clarifican la mecánica de las respuestas esperadas y concretan los objetivos que antes mencionábamos. Confeccionar ejemplos es costoso en tiempo y recursos al tiempo que necesario, ya que no todos los candidatos tienen por qué estar familiarizados con los distintos tipos de tareas existentes. Esto es algo que se suele ver con frecuencia en exámenes de idiomas destinados a la obtención de la nacionalidad o con fines migratorios, a los que pueden concurrir personas hablan un idioma perfectamente pero que jamás han hecho un examen. Finalmente, es recomendable que las instrucciones estén redactadas en la lengua meta y en un nivel inmediatamente inferior al de la prueba. Así, por ejemplo, las instrucciones de un examen de nivel C1 deberían estar redactadas en un lenguaje de nivel B2, las de un examen de B2 habrían de estar redactadas en un lenguaje de nivel B1, etc.

Como hemos dicho, los ítems de nuestras tareas están compuestos por una raíz, una clave y, en algunos casos, distractores. Por continuar con el ejemplo de tarea anterior y sin salir aún del aeropuerto, imaginemos que, una vez ubicados en nuestra puerta de embarque, escuchamos por el sistema de megafonía un anuncio que nos indica que se ha modificado el embarque. Podríamos convertir fácilmente esta tarea real en un ítem de respuesta múltiple:

iii ¿Cuál es la nueva puerta de embarque?
a) 15
b) 35
c) 45

En el ítem *iii* de más arriba, la raíz sería la pregunta «¿Cuál es la nueva puerta de embarque?». De entre las tres opciones que se ofrecen, la correcta es la que llamaremos clave y el resto las que llamaremos distractores. Como veremos a lo largo de las siguientes secciones, las raíces pueden adoptar formas muy variadas. Mientras que para cada ítem debe existir una única clave, el número de distractores puede variar. Por ejemplo, si el ítem tiene forma de pregunta abierta no será necesario facilitar distractores. Si se trata de un ítem de respuesta múltiple, lo habitual suele ser diseñar dos o tres distractores. Nuestra recomendación es diseñar una clave y dos distractores dada la evidencia (Abad *et al*. 2001; Jurado-Núñez, 2013; Gierl *et al*., 2017) de que los ítems así diseñados suelen poseer mejores propiedades psicométricas. Si el ítem está dentro de una tarea de emparejamiento, todas las opciones serán distractores para todos los ítems de esa misma tarea, etc. Los distractores que se diseñen han de estar extraídos del propio audio (i.e. contener palabras o ideas que se mencionen en la grabación), ser factibles, no descartables por sentido común y, sobre todo, no han de diseñarse con la idea de «tender trampas» al candidato, sino con la idea de comprobar si ha comprendido o no el contenido del texto o audio. Aunque no estrictamente necesario, es recomendable organizar la clave y los distractores de un ítem en orden alfabético, ya que con frecuencia las claves son lo primero que se diseña y, de no modificar su orden, estas tienden a ser siempre la primera opción.

Los ítems deberán aparecer en las tareas de forma secuencial con respecto al audio o texto al que corresponden. Es decir, el ítem sobre el contenido del minuto 1 debe aparecer antes que el ítem sobre el contenido del minuto 2, etc. En el caso de las tareas de comprensión auditiva, además, salvo casos muy concretos que requerirían de una justificación distinta, lo recomendable es permitir que el candidato pueda reubicarse en la tarea disponiendo de unos segundos que separen el momento en el que encuentra la respuesta a un ítem y el siguiente. Estos intervalos de tiempo regulares entre respuesta y respuesta pueden ser difíciles de encontrar en la grabación de discursos espontáneos. Es precisamente en estos casos en los que la guionización de la que hablábamos en la sección 3.4.2 puede ser útil, dado que nos permitirá espaciar adecuadamente el contenido importante e incluso incluir palabras que sirvan de baliza para señalar el final de un ítem y el comienzo de otro. No obstante, la irregularidad del discurso oral no invalida las grabaciones de monólogos o diálogos espontáneos que, para determinados constructos (en los que se asuman capacidades cognitivas avanzadas) pueden ser ideales.

A continuación, ilustramos algunos de los tipos de tareas más habituales en las pruebas de comprensión auditiva, dejando al margen aquellos modelos de tareas tradicionales centrados en la discriminación fonética, el acento, la entonación (*vid*. Heaton, 1979:58–65), el dictado u otros aspectos ortoépicos aislados que, a

nuestro parecer, han de ser evaluados en el marco general de la comprensión auditiva. Los ejemplos planteados en las siguientes secciones, como se verá, no son compartimentos estancos. En muchas ocasiones, pequeños matices en el diseño de una tarea pueden hacerla entrar en una u otra de las categorías descritas, que no pretenden ser un listado exhaustivo sino, más bien y de acuerdo con los principios que inspiran este libro, una guía práctica.

En los ejemplos que siguen identificaremos las tareas en función del tipo de ítems que contienen. La distinción más importante es la que separa las tareas con ítems de respuesta abierta de las tareas con ítems de respuesta cerrada. Los ítems de respuesta abierta son aquellos en los que el candidato tendrá que generar una respuesta *ex novo*. Los ítems de respuesta cerrada, por el contrario, son aquellos en los que se facilita al candidato una serie de respuestas predeterminadas de entre las que ha de elegir una. Las tareas con ítems de respuesta corta y reconstrucción son las únicas que se prestan exclusivamente al formato de respuesta abierta. Por otro lado, las tareas de respuesta múltiple y las de emparejamiento son las únicas que se prestan exclusivamente al formato de respuesta cerrada. Todas las demás, mediante determinados ajustes, pueden encajar tanto en un grupo como en otro.

La distinción entre ítems de respuesta abierta o cerrada es relevante en el diseño de la clave. La principal ventaja de los ítems de respuesta cerrada es la facilidad con que se pueden corregir. Si un ítem de este tipo está bien diseñado, solo tendrá una respuesta posible, lo que permite la corrección automática mediante sistemas de reconocimiento óptico si se realizan en papel, o mediante algoritmos si se realizan en soporte informático. Los ítems de respuesta abierta, sin embargo, han de pasar primero por el tamiz de la interpretación de un corrector o una inteligencia artificial, y esto puede dar lugar a variabilidad. El corrector, por ejemplo, ha de interpretar si una sintaxis pobre o una falta de ortografía al escribir determinada palabra son indicios de una falta de comprensión o de carencias de otro tipo. En este sentido, la recomendación es que al evaluar la comprensión auditiva se ha de dar prioridad a la comprensión propiamente dicha y no a la sintaxis o la ortografía en la medida en que estas no supongan un obstáculo para la inteligibilidad de la respuesta.

3.4.2.2. Respuesta múltiple

Las tareas con ítems de respuesta múltiple (*multiple-choice* o *multi-choice* en inglés) son, probablemente, las más comunes y las primeras que imaginamos al pensar en una prueba de dominio (Wolfe y Smith, 2007a:110). Los ítems de respuesta múltiple son aptos para medir distintos rangos de complejidad cognitiva. La raíz de estos ítems suele tener forma de oración enunciativa afirmativa incom-

pleta o de oración interrogativa directa o indirecta. La clave se presenta mezclada con los distractores para que el candidato la identifique.

A continuación, enumeramos un conjunto de recomendaciones para la redacción de tareas con ítems de respuesta múltiple que toma como base a Wolfe y Smith (2007a:112):

Recomendaciones sobre el contenido

1. Cada ítem debe reflejar un tipo de contenido específico.
2. Es recomendable mapear los textos (de audio o escritos) para consensuar el contenido sobre el que se preguntará (*vid.* 3.5.2).
3. Los ítems han de presentarse en el mismo orden en el que aparecen en el audio o el texto del que emanan.
4. Se puede parafrasear el contenido del audio e incluso utilizar palabras extraídas de este para la redacción de la raíz, clave y distractores del ítem.
5. Los ítems han de estar redactados en un nivel de lengua inferior al que intenta medir el examen.
6. Se han de evitar los ítems en los que tan solo hay una respuesta correcta y un distractor.
7. Se han de evitar los ítems cuya respuesta sea tan evidente que pueda identificarse sin necesidad de escuchar el audio o leer el texto correspondiente.
8. El contenido de un ítem ha de ser independiente del contenido del resto, es decir, la respuesta correcta de un ítem no ha de depender de la respuesta correcta de un ítem anterior.
9. Han de evitarse los ítems cuya respuesta pueda depender de la opinión del candidato.
10. En los niveles más altos de dominio se pueden usar ítems en los que se pregunte sobre la opinión de los participantes en la grabación o de los autores del texto escrito siempre que esta esté suficientemente clara.
11. Se han de evitar los ítems y distractores «trampa». El objetivo principal de un ítem es medir la capacidad de comprensión de los candidatos, no medir su habilidad para escapar de las trampas de un redactor de ítems.

Recomendaciones sobre el formato

12. Los ítems se deben presentar de forma vertical, no horizontal.
13. Los ítems de una misma tarea se deben presentar en una misma página del examen.
14. En el caso de las pruebas en soporte digital, los ítems que pertenezcan a un mismo audio han de estar en una misma pantalla. Es recomendable

analizar el comportamiento visual de los candidatos (*vid*. Bax, 2013 o Brunfaut, 2022).

15. En las pruebas de comprensión auditiva con soporte digital se ha de indicar si el candidato tiene control o no sobre el comienzo de la reproducción (cuya duración y progreso han de ser mostradas) y cómo puede ejercer dicho control.
16. En las pruebas en soporte digital el candidato debe poder escoger la respuesta de manera clara y poder corregir cualquier respuesta de una misma tarea antes de pasar a la siguiente tarea.

Recomendaciones sobre el estilo

17. Los ítems han de estar revisados (*vid*. 3.5.4, 3.5.5 y 3.5.6)
18. La gramática, puntuación, mayúsculas y ortografía han de ser correctas y coherentes con el manual de estilo de la institución.
19. Se debe utilizar lenguaje inclusivo en cuanto a género en las instrucciones y en el diseño de los ítems.
20. Se ha de minimizar la cantidad de lectura en cada ítem, por lo que si la clave y los distractores comienzan con la(s) misma(s) palabra(s), esta(s) ha(n) de ser incluida(s) en la raíz.

Recomendaciones sobre las instrucciones

21. Las instrucciones han de mostrarse claras y estar redactadas en un nivel de dominio inferior al del que se evalúa o, en caso de pruebas multinivel, en un nivel adecuado para el menor de los niveles evaluados.
22. Han de contextualizar la tarea para predisponer favorablemente al candidato a activar los esquemas mentales necesarios.

Recomendaciones sobre la raíz

23. La raíz ha de formularse de forma afirmativa evitando negaciones. En caso de que sea necesario incluir adverbios de negación (como «NO» o «EXCEPTO»), estos deben aparecer en mayúscula, no en cursiva.
24. La raíz también puede realizarse en forma de oración interrogativa directa o indirecta y, en la medida de lo posible, ha de evitarse la mezcla de estas formulaciones con las afirmativas.

Recomendaciones sobre la clave y los distractores

25. Las investigaciones indican que una clave y dos distractores son suficientes (Abad *et al*. 2001; Jurado-Núñez, 2013; Gierl *et al*., 2017). El diseño de un tercer y un cuarto distractor plausible no siempre es posible y estos suelen no ser funcionales (Tarrant *et al*. 2009).

26. Se ha de usar el mismo número de distractores en todas las tareas de un mismo componente y, si es posible, en todas las tareas de la prueba.
27. Se ha de comprobar que ninguno de los distractores puede ser también la clave.
28. La posición de la clave ha de ir cambiando entre los diferentes ítems para no generar un patrón predecible. Con frecuencia, la clave se redacta antes que los distractores y esto puede hacer que siempre quede en la primera posición. Para evitar esto, una vez terminada la redacción de las claves y los distractores estos se pueden reordenar alfabéticamente o de forma aleatoria.
29. Los distractores han de ser independientes y las soluciones no deben solaparse.
30. Las distintas opciones han de ser homogéneas en contenido, extensión y estructura gramatical.
31. Se desaconseja el uso de claves del tipo «ninguno de los anteriores». Estos distractores son difíciles de diseñar, ya que para que un distractor de este tipo sea válido las opciones anteriores han de descartarse como correctas de forma explícita en el audio o en el texto.
32. De igual manera, han de evitarse claves del tipo «todos los anteriores» o «no se menciona».
33. Al igual que la raíz del ítem, los distractores han de estar enunciados de manera afirmativa.
34. No se deben redactar intencionadamente distractores con ortografía o sintaxis incorrectas.
35. Ha de evitarse que los distractores sean similares entre sí, ya que esto puede conducir de forma automática a la elección de la clave.
36. Se han de evitar distractores absurdos o poco plausibles.
37. Para la redacción de los distractores se pueden usar respuestas típicas de candidatos en ítems similares que no contengan errores ortográficos o sintácticos.

En niveles de dominio bajos e intermedios, el uso de imágenes es particularmente interesante para presentar los distractores y la clave. Como evaluadores, las imágenes nos permiten observar de forma más directa la comprensión auditiva, ya que se eliminan posibles interferencias derivadas de carencias del candidato (sintácticas, problemas de comprensión lectora, etc.) que no están directamente relacionadas con la comprensión auditiva. Un ejemplo muy sencillo de este tipo de distractores sería aquel en el que, por ejemplo, se le presenta al candidato un audio en el que dos personas intentan ponerse de acuerdo sobre la hora a la que podrán verse. En el audio los intervinientes mencionan varias horas y luego se

deciden por una. Todas estas horas se presentan al candidato, por ejemplo, en forma de reloj para que escoja aquella por la que finalmente se deciden los hablantes de la grabación:

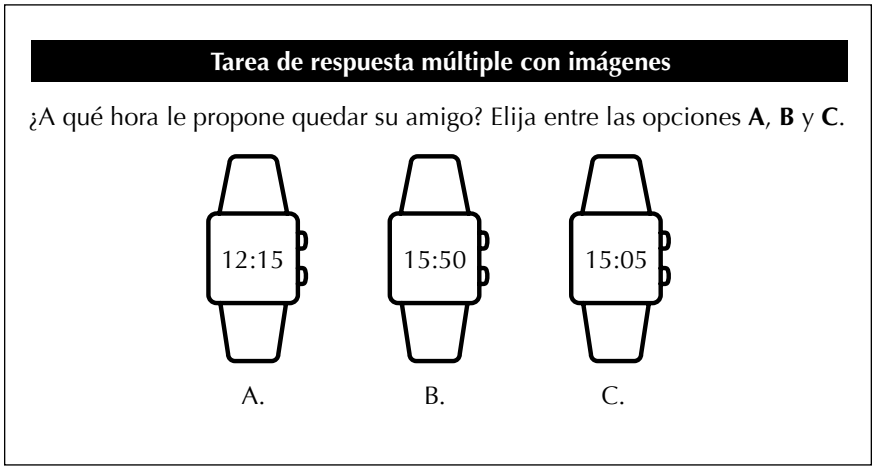


Figura 8. Tarea de respuesta múltiple con imágenes

3.4.2.3. Reconstrucción

Como su nombre indica, en este tipo de tareas el candidato ha de reconstruir una secuencia. Generalmente, se trata de secuencias temporales o procesos que se describen de forma lineal en el archivo de audio.

En estas tareas se puede ofrecer al candidato una serie desordenada de sintagmas, frases o imágenes para que las ordene según lo que escuche en la grabación. Es necesario incluir una opción extra. Si no se incluye una opción extra, cuando un candidato se equivoca en una respuesta está errando automáticamente en dos. Las imágenes son particularmente útiles en la evaluación de niveles inferiores como se aprecia en la figura 9.

Tarea de reconstrucción

A continuación va a escuchar a Yanet hablando de sus tareas diarias. Escuchará la grabación dos veces. Ordene las distintas actividades (**A-I**) de más abajo en los espacios correspondientes (**1-7**) según el orden en el que se realizan. Una de las actividades no es mencionada y, por tanto, no ha de usarse. Hemos incluido un ejemplo (**0**).

A.

B.

C.

D.

E.

F.

G.

H.

I.

¿En qué orden realiza Yanet sus actividades diarias?

0. C
1. ____
2. ____
3. ____
4. ____
5. ____
6. ____
7. ____

Figura 9. Tarea de reconstrucción

3.4.2.4. Emparejamiento

En este tipo de tareas (llamadas *matching* en inglés) los candidatos tienen que emparejar una serie de palabras, oraciones o párrafos con otros elementos.

Tarea de emparejamiento

Tras terminar su Grado en Ciencias Ambientales, usted ha decidido prepararse como Técnico Medioambiental para el Ministerio de su país. Durante el proceso de formación como Técnico usted participa en un coloquio sobre cambio climático.

Escuche la grabación y una las consecuencias del cambio climático que se enumeran más abajo (**1-8**) con los problemas a los que los vincula la oradora del coloquio (**A-D**).
Hemos preparado un ejemplo (**0**) para ayudarle.

~~**0.** Son un problema importante en Europa y el Mediterráneo.~~

1. Se sabe que causan estrés en la población.
2. Pueden tener impacto en las alergias.
3. En algunos casos son positivos/as.
4. Inutilizan los terrenos de cultivo.
5. Suponen pérdidas de hasta 150 billones de euros.
6. Pueden forzar la migración de la fauna local.
7. Son un problema para el suministro energético.
8. Mas de 3 millones de personas los sufren anualmente.

A. Olas de calor	**B.** Incendios forestales
0. Son un problema importante en Europa y el Mediterráneo.	
C. Inundaciones	**D.** Alteración de las estaciones

Figura 10. Tarea de emparejamiento

En la figura 10 de más arriba vemos que al candidato se le presentan ocho ítems (1–8) que ha de emparejar con cuatro ideas (A–D). En el caso de esta tarea en concreto, diferentes ítems pueden estar vinculados a una misma idea, aunque también puede diseñarse una tarea en la que cada ítem tenga que emparejarse con una sola idea como hemos visto en la figura 9. La elección sobre el diseño de la prueba, pues, afecta a las interacciones entre ítems, claves y distractores. Cuando cada ítem puede emparejarse con una sola idea, el resto de ideas son al mismo tiempo distractores para ese ítem y claves para otros. En estos casos es necesario redactar al menos una idea extra (i.e., que no haya de ser utilizada) para evitar que el candidato que se equivoca en un emparejamiento esté errando automáticamente en otro, como también hemos comentado al describir las tareas de reconstrucción. La presencia de esta idea extra, además, ha de indicarse en las instrucciones de la tarea.

Un ejemplo paradigmático de este tipo de tareas son las que se suelen encontrar en la parte 4 del componente de comprensión auditiva de los exámenes CAE y CPE de Cambridge University Press and Assessment (2024c:52; 55–56). Estas tareas están diseñadas para que los candidatos vinculen dieciséis ítems con cinco monólogos cortos (el equivalente a lo que hemos llamado ideas en el ejemplo de más arriba). La duración total de la grabación oscila entre los tres y los cuatro minutos y cada monólogo dura aproximadamente treinta y cinco segundos. La complejidad de esta tarea radica en el elevado número de ítems que se han de emparejar con los monólogos y en el hecho de que aquellos están, a su vez, divididos en dos ejes temáticos. Esto hace necesario que el candidato no solo comprenda los monólogos, sino también que posea una memoria de trabajo que le permita recordar todos los ítems que ha de emparejar. En nuestra opinión, este tipo de tareas son muy interesantes para los niveles superiores de dominio, si bien miden aspectos cognitivos que están más allá de la comprensión auditiva.

3.4.2.5. Respuesta corta

Este es un formato muy frecuente y el primero que nos viene a la cabeza cuando pensamos en ítems de respuesta abierta. Generalmente, la raíz de este tipo de ítems suele ser una oración interrogativa directa a la que el candidato ha de dar respuesta basándose en la información obtenida del registro sonoro.

Estas tareas son muy versátiles, puesto que permiten preguntar sobre ideas principales y secundarias, y posibilitan evaluar aspectos metalingüísticos, como por ejemplo, el registro o la actitud del autor frente al texto (*vid.* ítem 5 en la figura 11).

Si bien los ítems de respuesta corta son relativamente sencillos de confeccionar, su clave ha de ser definida con sumo cuidado. Es recomendable que en el enunciado de la tarea se establezca un número máximo de palabras para la res-

puesta y que se indique si las palabras que han de usarse en dicha respuesta deban proceder directamente del audio (sin modificaciones derivativas o flexivas) o si, por el contrario, el candidato podrá usar cualquier estructura de su repertorio, provenga esta o no de la grabación. En el caso de aquellas lenguas que lo permitan, se debe indicar si las contracciones se considerarán como una o dos palabras. Por último, se debe indicar si se pueden usar números o no en las respuestas y, en caso afirmativo, si estos se pueden escribir con letra o con número.

La clave, en la que se deben evitar nombres propios y palabras tabú, tendrá que ser una idea que pueda plasmarse en la extensión que determinen las instrucciones. Para ello, lo mejor es consensuar en el seno del equipo de confección qué respuestas se aceptarán y cuáles no en función de los resultados del mapeo (*vid*. 3.5.2). En ocasiones, los candidatos pueden ser sorprendentemente creativos al dar solución a estos ítems, por lo que su pilotaje es crucial (*vid*. 3.5.5). Pilotar nos ayudará a recopilar claves distintas a las establecidas inicialmente por los redactores que, en caso de ser adecuadas, habrán de pasar a formar parte de la clave definitiva de la tarea.

Tarea de respuesta corta

Tras terminar su Grado en Ciencias Ambientales, usted ha decidido prepararse como Técnico Medioambiental para el Ministerio de su país. Durante el proceso de formación como Técnico usted participa en un coloquio sobre cambio climático.

Escuche la grabación y responda con sus propias palabras a las preguntas de más abajo (**1-5**). Sus respuestas deberán contener un **máximo de 10 palabras**, incluyendo números (en caso de que sean necesarios), que podrán escribirse con letra o con número.

Hemos preparado un ejemplo (**0**) para ayudarle.

0. ¿Qué llevó al moderador del coloquio a interesarse por las olas de calor?

Que son un importante problema en Europa y el Mediterráneo.

1 ¿Qué problemas psicológicos puede causar el cambio climático en humanos?

_____

2. ¿Cuál es la consecuencia más perjudicial de las inundaciones?

_____

3. ¿Cuál es el impacto económico del cambio climático en Latinoamérica?

_____

4. ¿Qué tres propuestas hace el orador para frenar el cambio climático?

_____

5. ¿Cómo describiría la actitud del orador ante el cambio climático?

_____

Figura 11. Tarea de respuesta corta

También habrá de considerarse la importancia que tendrán la ortografía o la sintaxis en este tipo de tareas. No debe perderse de vista que el objetivo principal

de las tareas de comprensión auditiva de nuestros exámenes es valorar precisamente eso, la comprensión auditiva, y no necesariamente el dominio ortográfico o sintáctico de los candidatos, algo que se valorará en otra parte de la prueba. ¿Se debe considerar como válida una respuesta en la que el candidato parece demostrar haber entendido la información de la grabación, pero en cuya redacción comete errores ortográficos o gramaticales? La respuesta a esta pregunta dependerá de la interpretación que hagan los correctores del constructo de la prueba, una interpretación que puede diferir entre instituciones y que, como decíamos, puede estar ligada a los niveles evaluados. Dentro de un contexto académico, por ejemplo, el error ortográfico de un candidato de nivel C1 en una palabra de uso frecuente tendrá un peso distinto al que tendría el mismo error en un candidato de nivel A2.

3.4.2.6. Completar frases

Este tipo de tarea (*sentence completion* en inglés) contiene frases incompletas que parafrasean información del audio. La labor del candidato es comprender el significado de dichas frases y extraer del audio la información necesaria para dotarlas de sentido manteniendo una sintaxis correcta. Por sus características, este tipo de tareas suelen usarse cuando se desea evaluar la comprensión de detalles específicos.

Los ítems de este tipo de tareas comparten las ventajas y limitaciones de las tareas de respuesta corta que hemos visto en la sección anterior. Al igual que aquellos, son relativamente sencillos de diseñar, si bien su clave ha de ser definida con sumo cuidado. También aquí es recomendable que en el enunciado de la tarea se establezca un número máximo de palabras para las respuestas y si aquellas han de proceder del audio (sin modificaciones derivativas o flexivas) o si el candidato podrá usar cualquier estructura de su repertorio. Para determinar el número de palabras máximas se ha de tener en cuenta si las frases quedarán completas con palabras aisladas (contracciones, números, etc.) o si necesitan de sintagmas. También se habrá de considerar qué parte de la frase (inicial, intermedia, final) se facilita al candidato y si es necesario que este dote a las oraciones parafraseadas de coherencia sintáctica o si es suficiente con que demuestre haber entendido la información sobre la que se le pregunta. En las claves de estos ítems se han de evitar los nombres propios y las palabras tabú. De nuevo, el pilotaje de los ítems ayudará a observar si las expectativas de los diseñadores coinciden con lo que los candidatos suelen responder (*vid.* 3.5.5).

3.4.2.7. Rellenar huecos

Las tareas compuestas por ítems en los que se han de rellenar huecos (*fill in the blanks* en inglés) suelen estar enfocadas a evaluar si un candidato es capaz de comprender información específica.

De nuevo, los ítems de este tipo de tareas comparten las ventajas y limitaciones de los de respuesta corta y completar frases (*vid.* 3.4.2.5 y 3.4.2.6), y son muy similares a estos últimos. En los ítems para rellenar huecos, no obstante, es aconsejable que las palabras que el candidato tenga que extraer (evitando nombres propios y palabras tabú) se reduzcan al mínimo y que estas se extraigan directamente del audio sin modificaciones. Como suele ser habitual, cuanta más libertad se deje al candidato, más complejo será el diseño de la clave de la prueba. Siempre que se les pide a los candidatos que respondan con una o varias palabras extraídas directamente del audio se ha de vigilar que la sintaxis de las oraciones en las que estas hayan de encajar permita que se utilicen tal cual se escuchan. Así pues, si se busca que el candidato responda con un verbo concreto, este ha de poderse usar en las respuestas con la misma conjugación del audio; si se busca un sustantivo, este ha de poderse usar en el mismo género, número, caso, etc. con el que aparece en el audio.

Estas tareas se pueden cargar de realismo, por ejemplo, si se les da forma de apuntes tomados durante un discurso. En la tarea de la figura 12 las instrucciones delimitan claramente el número máximo de palabras que el candidato puede usar, indican que estas han de usarse sin modificar, y se especifica qué hacer con las cifras en caso de que formen parte de las respuestas. La indicación sobre la no modificación del audio es relevante, por ejemplo, para la clave del ítem 2, «pérdida de vidas humanas», que es lo que se escucha en la grabación. Si no se hubiese hecho explícita esta indicación y si justo antes del ítem 2 no se hubiese incluido el artículo «la», el candidato podría haber respondido, por ejemplo, «que se pierdan vidas», y habríamos de considerar su respuesta como correcta. La indicación sobre las cifras, por otro lado, es relevante para la clave del ítem 5, «150 billones de €/euros». Si no se hubiera indicado que las cifras han de expresarse con número, el candidato podría haber tenido problemas al considerar como respuesta «ciento cincuenta billones de euros», que suma cinco palabras y no las cuatro que se indican en las instrucciones. En el ítem 0, el ejemplo, la palabra «importante» está entre paréntesis para indicar al candidato que no es parte necesaria de la respuesta y que lo mismo podría ocurrir con los adjetivos de otros ítems, como por ejemplo, el 3, cuya clave es «(grandes) extensiones de cultivo». Esta tarea, además, incorpora la voz de varios hablantes de distinto género para ayudar al candidato a diferenciar entre el discurso de uno y otro. En el párrafo 2, por ejemplo, se indica que lo que sigue es parte de la intervención de una segunda oradora, algo que ayuda al candidato a ubicarse. Al candidato también le resultará particularmente útil encontrar en el texto de la tarea, entre ítem e ítem, «palabras baliza» extraídas directamente del audio. Esto ayuda al candidato a recuperar el hilo del discurso si pierde la concentración en algún momento. Gracias a estas palabras de referencia, será capaz de dejar atrás un hipotético ítem al que no haya sabido dar respuesta y reubicarse en la tarea para poder responder al siguiente. La separación del texto

en distintos bloques temáticos también es una interesante herramienta que nos permite estructurar conceptualmente la tarea y facilitar la audición al candidato.

Tarea de rellenar en los huecos (I)

Tras terminar su Grado en Ciencias Ambientales, usted ha decidido prepararse como Técnico Medioambiental para el Ministerio de su país. Durante el proceso de formación como Técnico, participa en un coloquio sobre cambio climático en el cual toma algunas notas.

Escuche la grabación y rellene los huecos de más abajo (**1-6**) con un **máximo de 4 palabras** que deberá extraer del audio **sin modificar**. Las cifras, en caso de que sean necesarias, se deben escribir con número. Hemos preparado un ejemplo (0) para ayudarle.

El moderador del coloquio comienza hablando de su experiencia personal sobre el cambio climático. Menciona que se interesó por este tema porque las olas de calor son ya un **0**. problema (importante)/(importante) problema en Europa y el Mediterráneo, lugar en el que vive. El coloquio continúa con la intervención de una experta en cambio climático de la República Dominicana que sugiere que el cambio climático ya está causando problemas emocionales en los seres humanos tales como **1**. _____. La experta hace referencia a las inundaciones que ha vivido en Haití, cuya consecuencia más devastadora es la **2**. _____ .
También se hace referencia a los incendios forestales, que inutilizan **3**._____. Los incendios, otra consecuencia del cambio climático, son particularmente dañinos para algunas comunidades selváticas centroafricanas, a las que aislan aún más puesto que destruyen tendidos eléctricos, lo que supone un problema para mantener el **4**._____ .
Por último, se menciona que la alteración de las estaciones ya está forzando la migración de algunas especies a zonas con climas más estables.
Según se estima, estas alteraciones y las consecuencias a ellas ligadas suponen ya pérdidas de **5**. _____ .
El moderador del debate interviene nuevamente al final y propone soluciones concretas contra el cambio climático como la descarbonización, la reforestación y el **6**._____.
Acaba con un mensaje esperanzador y optimista, animando a la audiencia a que todos contribuyan a la lucha contra el cambio climático.

Figura 12. Tarea de rellenar huecos (I)

Otra forma muy interesante de preparar este tipo de tareas consiste en usar mapas o gráficos. Los gráficos se suelen adaptar particularmente bien a los archivos de audio que describen procesos, mientras que los mapas pueden usarse, por ejemplo, con la previsión del tiempo meteorológico que extraeremos fácilmente de boletines informativos radiofónicos. También se puede utilizar el mapa más detallado de, por ejemplo, el campus de una universidad, en el que se hayan eliminado los nombres de determinados edificios (biblioteca, cafetería, aularios, gimnasio, etc.) para que el candidato identifique cuál es cuál tras escuchar una conversación en la que alguien describe el campus.

Como ya hemos comentado anteriormente en varias ocasiones, el uso de materiales visuales (fotos, mapas o diagramas) suele estar ligado a la evaluación de niveles bajos de dominio, puesto que hacen que la evaluación dependa en menor medida de aspectos como la gramática, el vocabulario o la capacidad lectora (Heaton, 1979:65), como ocurre en las primeras tareas de comprensión auditiva de los exámenes A2 del DELE del Instituto Cervantes (2022).

Es posible utilizar esta información gráfica también en niveles altos de competencia con tareas ligadas a formas más elaboradas de constructo. Por ejemplo, podríamos diseñar una tarea en la que nuestros candidatos tuvieran que interpretar un gráfico de barras o un mapa conceptual, completar el esquema de un proceso de fabricación, etc. Este tipo de tareas podrían considerarse como una categoría distinta, a caballo entre las tareas de rellenar huecos y las de emparejamiento que, además, abren la puerta a otros tipos de evaluación. Distintas escuelas oficiales de idiomas españolas, por ejemplo, han utilizado información gráfica compleja integrada con comprensión auditiva para el diseño de tareas orientadas a medir la mediación, en las que se pide al candidato que interprete y simplifique determinado contenido para transmitirlo de forma más eficiente a hablantes de niveles inferiores.

Vivimos en un mundo cada vez más dominado por la imagen en el que este tipo de tareas se dan con mucha frecuencia en la vida real. Los manuales de montaje de muebles, los manuales de instalación de un dispositivo electrónico o los mapas son solo algunos ejemplos. El color ha sido una consideración importante en los exámenes tradicionales en papel. Si el candidato necesita los colores para la interpretación del grafismo, hemos de asegurarnos de que tenemos capacidad para que las copias del examen estén en color. Dado que esto puede suponer un coste adicional, se suele intentar diseñar tareas en las que la imagen esté en blanco y negro. Obviamente, esto no será un problema para las pruebas realizadas en soporte digital.

Tarea de rellenar huecos (II)

Usted está planeando un viaje a Colombia y antes de partir decide escuchar un boletín informativo para conocer el clima que tendrá en las distintas ciudades que visitará.

Escuche la grabación y rellene los huecos de más abajo (**1-6**) con **una sola palabra o número**, que deberá extraer del audio **sin modificar**. Hemos preparado un ejemplo (**0**) para ayudarle.

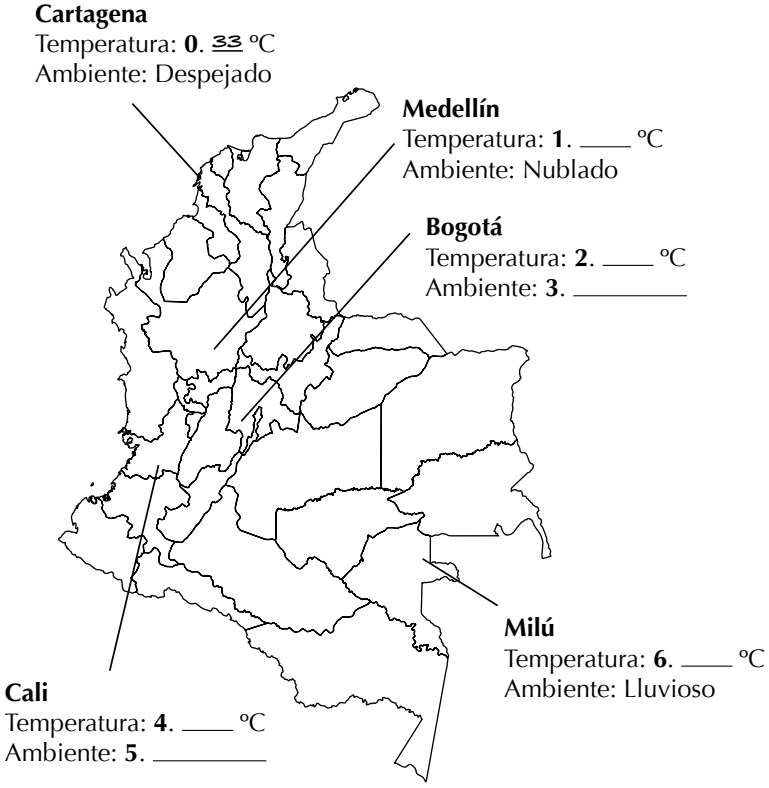


Figura 13. Tarea de rellenar huecos (II)

3.4.2.8. Tareas de destrezas integradas

Una tarea de destrezas integradas puede tener cualquiera de los formatos descritos en las secciones anteriores. Si las incluimos en este punto es porque traen consigo una serie de consideraciones que los redactores han de tener en cuenta durante el diseño de sus pruebas.

Como su nombre indica, las tareas de destrezas integradas son aquellas en las que el candidato ha de manejar dos o más destrezas para encontrar las respuestas correctas. En sentido estricto, cualquier tarea de comprensión auditiva es integrada, ya que en algún momento el candidato ha de leer las instrucciones de la tarea y redactar su respuesta. Lo que diferencia a una tarea tradicional de una tarea con destrezas integradas no es, pues, la interacción de destrezas, sino el peso que estas tengan y el uso que de ellas haga el candidato, por ejemplo, al tener que examinar varias fuentes con el objetivo de seleccionar ideas para sintetizarlas transformando el lenguaje original según convenciones estilísticas (Knoch y Sitajalabhorn, 2013:306).

La idea de evaluar una lengua como un todo en el que confluyen distintas destrezas y no como una serie de habilidades aisladas no es nueva. Esta interpretación ha sido potenciada por el lugar cada vez más relevante que las tareas han adquirido en la enseñanza, por el auge del aprendizaje integrado de contenidos y lenguas (AICLE) y por la formación en contextos de inmersión (Plakans, 2012:249). Desde la década de los setenta del siglo XX han existido distintas corrientes que han abogado por una visión unitaria de las destrezas en la evaluación de lenguas. Los primeros planteamientos teóricos partieron de la llamada gramática expectativa (*expectancy grammar* en inglés), según la cual el dominio de una lengua se construye mediante la interacción de elementos lingüísticos y extralingüísticos. Una interacción correcta entre ambos elementos permitiría al hablante predecir determinadas formas lingüísticas. Los *cloze tests* (tareas de comprensión lectora en las que se elimina una de cada x palabras, de las que hablaremos en la sección 3.4.3.4) serían un ejemplo de tareas encuadradas en esta concepción. La lectura de la tarea (elementos lingüísticos) introduciría al candidato en el contexto de la tarea (elementos extralingüísticos), y fruto de la interacción entre ambos el candidato podría predecir las palabras que han sido eliminadas. Un segundo conjunto de aproximaciones teóricas se centró en la autenticidad de las destrezas integradas, entre ellos, como Brunfaut (2016:100–101) indica, los postulados comunicativos que, en su búsqueda de la autenticidad y el realismo, consideran que las tareas integradas reflejan mejor la vida real. La tercera corriente de pensamiento sobre las destrezas integradas procede de la literatura sobre L1 y la psicología educativa y cognitiva. Esta corriente defiende que las destrezas lingüísticas comparten determinados procesos cognitivos y que ello justificaría su evaluación de una manera integrada. Así, por ejemplo, dado que la escritura puede influir en el desarrollo de la lectura y viceversa, de nuevo tendría

sentido evaluarlas de forma integrada (Plakans, 2012:250–251). En la actualidad, las tareas que integran la comprensión auditiva con otras destrezas se utilizan en exámenes como TOEFL iBT o en los niveles C de los exámenes DELE, en los que los candidatos han de resumir por escrito una conversación o una lección.

Los mismos argumentos que justifican la presencia de tareas integradas en los exámenes de dominio nos plantean cuestiones adicionales que, a su vez, vienen acompañadas de nuevos retos. Desde el punto de vista biolingüístico, por ejemplo, es importante determinar si existen diferencias en los procesos cognitivos de los candidatos al afrontar las tareas de uno y otro tipo (Plakans, 2008), o en la forma en que estos activan estrategias (Swain *et al*. 2009). Las relaciones entre estos procesos cognitivos, los umbrales de activación de estrategias y las jerarquías entre destrezas podrían responder al interrogante que Plakans (2012:254) plantea acerca de si la puntuación de un candidato en una tarea de destrezas integradas refleja realmente su competencia en todas las destrezas en liza o solo en alguna de ellas. Como vemos, está por determinar en qué medida el realismo de estas tareas va de la mano de evidencias que lo justifiquen.

En cualquier caso, aquellos que confíen en las tareas de destrezas integradas habrán de tener particular cuidado en sus diseños. La elección de un formato de tarea adecuado es capital a la hora de integrar destrezas. Si no somos capaces de facilitar al candidato un *input* suficiente y adecuado, de establecer unos objetivos claros cuyo *output* pueda ser calificado adecuadamente a través de una escala de corrección en la que no haya sobrerrepresentación del constructo, estaremos simplemente creando una tarea convencional revestida de un falso halo de integración. En la figura 14 mostramos un ejemplo de tarea en la que se integran comprensión auditiva y producción escrita en un examen realizado en soporte digital. Navarrete (2022) propone distintas e interesantes tareas que integran la comprensión auditiva no solo con la producción oral y escrita sino también con aspectos de mediación, yendo así más allá de la concepción tradicional de destrezas y acercándose al esquema de modos de comunicación que propone el MCER (Consejo de Europa, 2001) (*vid*. figura 7). Ejemplos menos sofisticados, pero igualmente interesantes, son los antes mencionados de los exámenes TOEFL iBT o DELE.

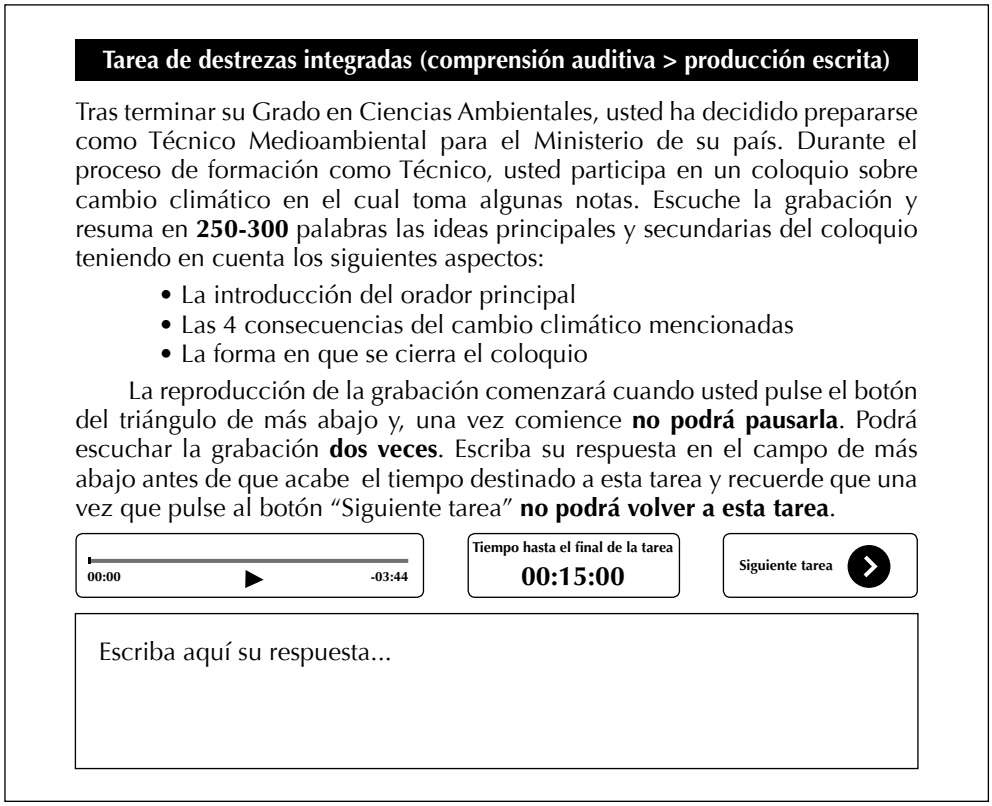


Figura 14. Tarea de destrezas integradas

3.4.3. Comprensión lectora

La lectura es el envés de la escritura y, al igual que esta, un artificio humano. Una lengua puede existir sin ser escrita o leída, pero no sin ser hablada. Al nacer escuchamos, después hablamos y solo cuando nos sumergimos en determinados contextos culturales aprendemos a leer y a escribir.

A lo largo de la historia, la lectura no siempre se ha practicado de la misma manera. Vallejo (2020:61) nos cuenta cómo hasta la Edad Media «la norma era leer en voz alta, para uno mismo o para otros» y que «los escritores pronunciaban las frases a medida que las escribían escuchando así su musicalidad. Los libros no eran una canción que se cantaba con la mente, como ahora, sino una melodía que saltaba a los labios y sonaba en voz alta». Vallejo (*ibid.*) continúa diciendo que, tal vez por esa razón, los primeros en leer en silencio, «en conversación muda con el escritor, llamaron poderosamente la atención», y que en el siglo IV, San Agustín de

Hipona «se quedó tan intrigado al ver leer de esta forma al obispo Ambrosio de Milán, que lo anotó en sus *Confesiones*».

Una actitud silente ante el texto escrito también nos permite leer de forma no secuencial. Si comparamos la manera en que leemos con la manera en que escuchamos nos daremos cuenta de que al escuchar las palabras nos llegan incesantemente, una tras otra y, salvo que el hablante repita intencionadamente parte de su alocución, no está en manos del oyente volver atrás en la sucesión de sonidos que componen el discurso oral. En el caso de la lectura, sin embargo, tenemos en todo momento la posibilidad de volver hacia atrás o avanzar, de repetir lo ya leído, de dejar huecos con respecto a lo que vendrá. Técnicas de análisis como el seguimiento del movimiento ocular (*eye-tracking* en inglés) demuestran algo que ya se sospechaba (Goodman, 1967:131), que la lectura de un texto en un examen no tiene siempre lugar de izquierda a derecha y de arriba abajo. Más bien al contrario, según ha demostrado gráficamente Brunfaut (2022), los candidatos leen obviando algunas partes del texto y centrándose en otras, entre las que se encuentran los propios ítems, que no son parte del texto, pero sí una parte fundamental de la tarea (Bax, 2013). Brunfaut (2022) muestra mediante gráficas de nódulos la secuencia en que los candidatos a una prueba leen. Los diferentes nódulos están organizados por números según el orden en el que el candidato los recorre, y son de mayor o menor tamaño en función del tiempo que este se detiene en ellos. Estos análisis de movimiento ocular no dejan ninguna duda de que, como decíamos, la lectura no es un proceso lineal.

El análisis científico del fenómeno de la lectura, al igual que la forma en que leemos, también ha variado a lo largo del tiempo. Hacia mitad del siglo XX, por ejemplo, se intentó explicar el funcionamiento de la lectura a través del conocido como modelo de procesamiento desde arriba (*top-down* en inglés), (Goodman, 1967) que presentaba la lectura como un juego de adivinación (*guessing game*), un proceso psicolingüístico en el que el conocimiento previo del lector tendría un papel fundamental. Más tarde, hacia finales del siglo XX surgió el modelo de procesamiento desde abajo (*bottom-up* en inglés) (Davoudi y Hashemi, 2015:174). Esta concepción de la lectura postula que el lector activaría determinados esquemas de conocimiento mediante la recombinación de unidades lingüísticas de nivel inferior con otras de complejidad cada vez mayor. Así, a través de la unión de letras el lector llega a la sílaba, mediante la combinación de sílabas a la palabra, de esta se pasa al sintagma, etc. Otros modelos son el interactivo (Rumelhart, 2013:172), que explica la lectura como el proceso de interacción entre un repositorio visual de información y un sistema de extracción capaces de alimentar un sintetizador que genera la interpretación más plausible de la información recibida; el interactivo compensatorio (Davoudi y Hashemi, 2015:175), según el cual un lector que experimenta problemas en una dimensión de procesamiento es capaz de compensarlo con sus destrezas en otra dimensión lectora; el modelo situacional

(*ibid.*175–178); el modelo de LaBerge-Samuels (Rumelhart, 2013:720); el modelo de Gough (*ibid.*), etc. Para una discusión más detallada de estos y otros modelos puede consultarse Alvermann *et al.* (2013).

Hallazgos sobre los patrones de movimiento ocular durante la lectura como los descritos más arriba nos conducen en la actualidad a enfoques empíricos vinculados de forma directa con las ciencias naturales (Chomsky, 2000:106). Desde un enfoque biolingüístico, se están dando pasos en la identificación de los factores genéticos y ambientales que influyen en la comprensión lectora. Keenan *et al.* (2008; 2009), al investigar estos dos factores en la capacidad lectora en L1 de gemelos idénticos y no idénticos han llegado a la conclusión de que diferentes exámenes de comprensión lectora miden distintas destrezas y que, incluso, el mismo test puede medir cosas diferentes en función de la edad del candidato. Por ejemplo, descubrieron que la capacidad de vincular grafemas con sonidos (*decoding* en inglés) tiene un impacto mayor en ítems cortos, en los que el éxito parece estar vinculado a la correcta asociación de grafía y sonido en una única palabra (Keenan *et al.*, 2009:235).

Gracias a estos enfoques sabemos que la habilidad lectora es, al igual que la obesidad, la salud cardiovascular o el consumo de drogas, el resultado socialmente significativo de complejas interacciones entre factores genéticos, biológicos y ambientales (Petrill, 2009:258). Uniendo todos estos descubrimientos y otros como el de Tang *et al.* (2023), que logró hacer saltar el pensamiento directamente del cerebro humano al disco duro de un ordenador, tal vez en algún momento seamos capaces de evaluar la comprensión lectora de nuestros candidatos mediante resonancia magnética funcional o métodos similares. Aunque son pocas las certezas que tenemos sobre los fundamentos biológicos de la comprensión lectora, lo que sí sospechamos es que los futuros avances vendrán de la mano de la interacción entre las disciplinas que estudian las diferencias entre individuos (Buscher *et al.* 2010), las que analizan el desarrollo individual del lector y la relación de este con una compleja y cambiante realidad digital (Korte, 2020), y las disciplinas centradas en la genética conductual o en aspectos neurocognitivos (Wagner *et al.*, 2009), entre otros.

Todos los estudios sobre comprensión lectora mencionados en esta sección han requerido datos que, a su vez, se consiguen mediante las respuestas de los candidatos a determinadas tareas. Lo que ahora sigue es la descripción de distintos formatos de tareas destinadas, precisamente, a recabar dichos datos con los que seguramente en el futuro llegaremos a entender mejor los fundamentos biológicos de la comprensión lectora del ser humano. Esta enumeración de formatos, al igual que ocurría en el caso de la comprensión auditiva, no pretende ser exhaustiva (para eso puede consultarse Alderson, 2000:202–270), sino más bien práctica. Aunque es posible que el tiempo nos enseñe nuevas formas de medir la comprensión lectora, las que ahora conocemos y las que se usan con más frecuencia son las que siguen.

3.4.3.1. Respuesta múltiple

En general, los ítems de respuesta múltiple en comprensión lectora comparten las mismas características y recomendaciones sobre formato, estilo, raíz y distractores que los ítems de respuesta múltiple de comprensión auditiva descritos en la sección 3.4.2.2, por lo que recomendamos revisar dicha sección.

Igual que en el caso de las preguntas de respuesta múltiple para comprensión auditiva, el redactor de ítems habrá de determinar si lo que desea medir con estos es la comprensión general de un candidato o su habilidad para identificar información específica.

3.4.3.2. Reconstrucción

Dentro de esta sección incluiremos dos subtipos de tareas. En primer lugar, tenemos aquellas que son idénticas a las de comprensión auditiva descritas en la sección 3.4.2.3, es decir, tareas en las que se pide a un candidato que ordene una serie de párrafos para establecer una secuencia que viene dada por un texto referente.

En segundo lugar, tenemos un tipo de tarea que, debido a su aspecto visual, puede ser considerada como una tarea de emparejamiento. Nosotros, sin embargo, preferimos considerar a este subtipo como tareas de reconstrucción, ya que lo que se le pide al candidato es que reconstruya la estructura lógica de un texto en función de la coherencia interna de sus párrafos. Para diseñar este tipo de tareas (que en inglés se conocen como *gapped text*, o texto con huecos) lo habitual es seleccionar un texto auténtico, extraer varias frases o párrafos del mismo y pedirle al candidato que los reubique en el lugar que corresponda, tal y como se observa en la tarea de la figura 15 de más abajo.

Tarea de reconstrucción *(gapped text)*

Ha decidido programar un viaje a Colombia e investigando sobre las tradiciones musicales del país se encuentra con el texto de más abjo, que lleva por título *La cumbia: pasado, presente y futuro,* del que han sido extraidos 4 párrafos.

Coloque los párrafos (**A-G**) en el lugar del texto principal que corresponda (**1-5**). Observe que hay un párrafo adicional que no necesita utilizar. Hemos preparado un ejemplo (**0**) para ayudarle.

Donde suena la cumbia, aparecen sonrisas. Este género musical que representa a Colombia en el mundo es un fiel testimonio de la riqueza rítmica y cultural del país. En este artículo haremos un recorrido por la historia de este ritmo caribeño. Comenzaremos por la etimología de la propia palabra «cumbia» y más tarde veremos cómo esta música, caribeña en origen, se expande y se adapta a distintos contextos culturales.

0. Párrafo A

La experta en música costeña colombiana, María del Pilar Jiménez González, sitúa su origen alrededor del siglo XVIII, en la costa atlántica de Colombia, y describe su formación como el resultado de un largo proceso de fusión de tres elementos etnoculturales como son los indígenas, los blancos y los africanos, de los que adopta las gaitas, las maracas y los tambores.

1.

En el ciclo de formación de la cumbia, María del Pilar Jiménez resalta los tiempos de Simón Bolívar (1800), época durante la cual el alegre ritmo caribeño se fue fraguando en la parte alta del valle del río Magdalena.

2.

Después de su nacimiento, la mayor transformación de la cumbia se registra en los años treinta del siglo XX. Las clases acomodadas se apropian del género, de su ritmo y expresiones populares, y esto obliga a un cambio estético en una música eminentemente instrumental que hasta ese momento había sido menospreciada. A partir de ese momento, la cumbia dejó de ser exclusivamente instrumental, pasó a tener letras y evolucionó integrando otros instrumentos, como el acordeón o, más recientemente, instrumentos electrónicos y orquestación completa, lo que la ha hecho accesible a un público más diverso no solo en Colombia sino también en el resto de Latinoamérica.

3.

Una de las más famosas variaciones de cumbia es la llamada cumbia rebajada. La historia dice que un día en Monterrey, Méjico, el DJ Gabriel Duéñez estaba poniendo música en una fiesta muy larga que duró más de cinco horas. En ese momento, por falta de energía (se habían recalentado los equipos), todo comenzó a sonar de manera diferente.

4.

¿Qué es lo que estaba sonando? Cumbia colombiana con elementos del vallenato y el toque sabanero. Así nació la famosa cumbia rebajada, cuyo origen, características y evolución ahora conocemos mejor.

5.

Párrafos extraídos

~~**A.** La raíz de la palabra «cumbia» parece proceder del vocablo africano «cumbé», que significa jolgorio o fiesta. La cumbia es por lo tanto la imagen viva de la fiesta y de la influencia africana en el caribe y, cada vez más, en el resto del mundo.~~

B. Por lo tanto, la cumbia colombiana no es la única del continente, pues desde mediados del siglo XX se difundió por buena parte de Latinoamérica, dejando variaciones del género como la cumbia argentina, mexicana, salvadoreña y peruana.

C. Los discos empezaron a pasarse más lentos y las canciones se sentían mas pastosas y suaves: rebajadas. Gabriel intentó reparar el equipo pero como la gente siguió bailando lo dejó así.

D. Estos son los mismos ingredientes que se aprecian en gran parte de la historia latinoamericana, donde confluyen la herencia de los pueblos precolombinos y la mano de obra esclava traída desde África por los españoles para resolver el problema que se imponía en las plantaciones. De la mezcla de los tambores africanos y la romanza española, nace la cumbia.

E. Más allá de estas fechas y esta ubicación aproximadas, poco se conoce sobre el lugar exacto del nacimiento de la cumbia. Y es que, tal vez desde su nacimiento, la cumbia haya estado destinada a ser una música viajera, imposible de amarrar con coordenadas geográficas. Como bien dice la canción Yo me llamo cumbia: «Yo nací en las bellas playas Caribes de mi país. Soy barranquillera, cartagenera, yo soy de ahí. Soy de Santa Marta, soy monteriana pero eso sí, ¡yo soy colombiana, o tierra hermosa donde nací!».

F. Muy probablemente serán fusiones musicales como las de la cumbia rebajada las que permitan a la cumbia seguir evolucionando para expandirse más allá de Latinoamérica para, tal vez, conquistar muy pronto el resto del mundo.

G. Se denomina cumbiamba al conjunto de parejas de bailarines acompañado del conjunto que ejecuta la cumbia. Mientras que la cumbia se toca con banda, y las bailarinas llevan velas o teas en las manos, la cumbiamba se baila con acordeón y flauta de millo y sin velas.

Figura 15. Tarea de reconstrucción

Al diseñar este tipo de tareas ha de tenerse en cuenta que, al igual que ocurrirá con las tareas de emparejamiento que describiremos en la sección 3.4.3.3, no basta con extraer frases o párrafos de forma aleatoria. Las secciones extraídas han de poder ser ubicadas inequívocamente en uno y solo uno de los huecos del texto referente. Para asegurarnos de que esto es así es recomendable extraer secciones que contengan suficientes referencias anafóricas (i.e. a elementos mencionados anteriormente) y catafóricas (i.e. referencias a elementos que aparecerán con posterioridad). Los pronombres y los antecedentes de oraciones de relativo son buenos referentes anafóricos. Las conjunciones también pueden ayudar al candidato a establecer el lugar exacto en el que ha de ubicarse un párrafo.

Por ejemplo, en la tarea de la figura 15 vemos que el párrafo que encaja en el hueco 0, el ejemplo, es A porque en el primer párrafo se menciona que se co-

menzará hablando de la etimología de la palabra «cumbia». En el hueco 2 la única opción posible es la D, puesto que en el segundo párrafo se habla de «indígenas, blancos y africanos», que es precisamente lo que desarrolla D. El resto de emparejamientos son 3-B, 4-C y 5-F. El párrafo G no se usa porque alude a una variante de cumbia tradicional que no se menciona en ninguna parte del texto dado

3.4.3.3. Emparejamiento

De nuevo, las tareas de emparejamiento en comprensión lectora comparten las características de las de comprensión auditiva, por lo que en este punto es recomendable revisar la sección 3.4.2.4.

Un ejemplo de tareas de emparejamiento de comprensión lectora es aquel en el que se da al candidato un texto sobre varios conceptos y se le pide que vincule una serie de aseveraciones con cada uno de dichos conceptos. Por ejemplo, se le pueden ofrecer al candidato tres textos que hablen sobre el turismo en Santiago de Chile, Buenos Aires y Madrid y se le pide que empareje una serie de aseveraciones con la ciudad que corresponda.

Otra forma bastante habitual de emparejamiento es aquella en la que se usa un texto con diferentes secciones, cada una de las cuales está precedida por un encabezado que la resume. Al diseñar la tarea, estos encabezados se extraen del texto original y se pide a los candidatos que los reubiquen en el lugar que les corresponde. En esta forma de emparejamiento, si se usan textos auténticos, hemos de tener en cuenta que, obviamente, el autor original de los mismos no los escribió pensando que debían ser resúmenes perfectos de las secciones que siguen, porque algún día nosotros los usaríamos en una tarea de comprensión lectora. Puede que, en sus encabezados, el autor original dirigiese la atención a una parte específica de la sección, o incluso que redactase un encabezado que pudiera encajar en varias secciones. Por lo tanto, al diseñar tareas de emparejamiento mediante encabezados es importante cerciorarse de que estos encajan en una y solo una de las secciones. Si los encabezados del texto original no cumplen con estas características han de modificarse para que la vinculación a cada una de las distintas secciones sea inequívoca, y poder medir así de forma fidedigna la capacidad de comprensión lectora de nuestros candidatos.

3.4.3.4. *Cloze*

Este tipo de tareas está en desuso. En su forma básica se diseñan extrayendo palabras de un texto a intervalos regulares (por ejemplo, una de cada cinco, de cada diez, etc.). Se suelen dejar algunas frases completas tanto al comienzo como al final del texto con el objetivo de contextualizar la lectura. La labor del candidato en estas tareas es encontrar las palabras que han sido extraídas (*vid*. Alderson, 2000:207–211).

La principal ventaja de este tipo de tareas es que son relativamente sencillas de diseñar. Es suficiente con encontrar un texto que se adapte a nuestras necesidades y aplicarle un intervalo de eliminación. Después de aplicado, se ha de comprobar que las palabras extraídas admiten únicamente una solución. Si son varias las palabras que pueden encajar en el hueco abierto, se ha de plantear la posibilidad ampliar la clave, es decir, de aceptar como válidas todas las palabras que encajen sintáctica y semánticamente en la frase, lo que puede suponer complicaciones durante la corrección de las pruebas. Para evitar este problema puede diseñarse una variante, el llamado *closed cloze* (o *cloze* cerrado), en el que se le dan al candidato las palabras extraídas mezcladas con otras que no encajan en los huecos para que seleccione las correctas.

El principal motivo que ha llevado a estas tareas a estar en desuso es el hecho de que parecen apelar más a la competencia léxica y gramatical de los candidatos que a su capacidad de entender un texto, motivo por el cual suelen estar relegadas a la medición del uso de la lengua en aquellas pruebas que incluyen este componente.

3.4.3.5. Respuesta corta

Al igual que en el caso de comprensión auditiva, este es un formato muy frecuente en la evaluación de comprensión lectora y el primero que nos viene a la cabeza cuando pensamos en ítems de respuesta abierta.

Las respuestas a estos ítems no están sujetas a la aleatoriedad de otros ítems como, por ejemplo, los de respuesta múltiple, en los que el candidato puede llegar a elegir la respuesta correcta por azar. Dado que no es necesario redactar distractores, su diseño es relativamente sencillo. No obstante, esta sencillez tiene la contrapartida de la complejidad de la clave, ya que los candidatos pueden dar multitud de respuestas correctas diferentes. Las tareas de comprensión lectora de este tipo comparten las características de las de la sección 3.4.2.5.

3.4.3.6. Completar frases

De nuevo, las tareas con este tipo de ítems comparten las características de las de comprensión auditiva descritas en la sección 3.4.2.6 que, a su vez, comparten características con los ítems de respuesta corta tanto de comprensión auditiva (3.4.2.5) como de comprensión lectora (3.4.3.5).

3.4.3.7. Rellenar huecos

Estas tareas son similares a las descritas en la sección 3.4.2.7 para comprensión auditiva. Al igual que aquellas, en el caso de la comprensión lectora también es frecuente el uso de imágenes (gráficos, mapas, etc.).

3.4.3.8. Verdadero/falso con justificación

Este tipo de ítems, exclusivos de la comprensión lectora, permiten evaluar tanto la comprensión de información general como específica, si bien suelen adaptarse mejor a esta última.

Para el diseño de este tipo de ítems se redacta una aseveración sobre el texto elegido para la tarea y se pide al candidato que indique si dicha aseveración es verdadera o falsa. Con el objeto de evitar que los candidatos respondan aleatoriamente, algo que distorsionaría sus calificaciones, debemos solicitar la justificación de la respuesta. Esta justificación puede ser de varios tipos. Se puede solicitar a) la redacción de una frase corta con sentido, b) que se indique en qué línea del texto se encuentra la frase que justifica la respuesta o c) se puede pedir al candidato que copie la frase que contiene la información que le ha llevado a marcar la aseveración como correcta o incorrecta. La opción a) tiene la ventaja de que nos podría permitir evaluar no solo la comprensión lectora, sino también la producción escrita del candidato (si eso es lo que deseamos), y comparte las desventajas de cualquier tipo de respuesta abierta, es decir, la multitud de formulaciones posibles y las complicaciones que esto puede suponer en la corrección. La opción b) requiere que cada línea del texto esté debidamente numerada y que la justificación a la respuesta se concentre en una sola línea, lo que pude suponer complicaciones a la hora de dar formato a la tarea. Desde nuestra experiencia, la opción c) es la que ofrece mejores posibilidades si se articula de forma adecuada. En estos casos, lo más funcional es pedir a los candidatos que copien las tres o cuatro primeras palabras de la oración que justifique su respuesta. En ocasiones se solicita al candidato que copie la frase o el sintagma que contiene la justificación, algo que acarrea problemas en la corrección. Por un lado, copiar una frase completa puede requerir tiempo y espacio en el examen. Por otro, los candidatos no tienen por qué saber qué es un sintagma o cómo acotarlo y puede que la justificación se colija por la interrelación de varios de ellos. Hemos de asegurarnos de que la información solicitada, de nuevo, se encuentra solo en una frase, aquella cuyas tres o cuatro primeras palabras ha de copiar el candidato. Sea cual sea la formulación elegida, las instrucciones de la tarea han de ser claras. En este tipo de ítems, una respuesta sólo ha de considerarse como válida si el candidato indica correctamente si la aseveración es verdadera o falsa y la acompaña de la justificación adecuada. Si una de ambas partes falla, las respuestas han de considerarse incorrectas.

Con frecuencia hemos encontrado que los candidatos que se enfrentan por primera vez a este tipo de tareas responden sin tener totalmente en cuenta las instrucciones de la tarea. En ocasiones responden con una explicación que no es pertinente, señalan varias líneas del texto en su justificación o copian más palabras de las solicitadas. En este punto es importante recordar que los candidatos se

enfrentan a una prueba de comprensión lectora y que, si las instrucciones están debidamente redactadas, entenderlas y seguirlas es también un indicador de comprensión lectora.

3.4.3.9. Tareas de destrezas integradas

A la hora de diseñar tareas de destrezas integradas en la comprensión lectora se han de tener en cuenta los mismos aspectos sobre el equilibrio del peso de las distintas destrezas en liza que ya hemos descrito en la sección 3.4.2.8.

Los formatos en los que con más frecuencia se suele integrar la comprensión lectora son aquellos que van de la lectura a la producción escrita u oral. Por ejemplo, se ofrece al candidato un texto sobre un tema específico y más tarde se le pide que realice un resumen del mismo, bien sea escrito u oral.

3.4.4. Producción e interacción escritas

Al hablar, las palabras desaparecen justo después de ser pronunciadas. El ser humano ha creado la escritura para capturar las lenguas, para hacerlas perdurar. *Verba volant, scripta manent*.

La escritura es, sin duda, una de las más importantes revoluciones tecnológicas de la historia. Lo es porque, entre otras cosas, ha potenciado la cooperación humana y hecho posible el registro de datos y eventos que de otra manera habrían quedado en el olvido. Lo que se conoce del mundo antes de la aparición de la escritura es mucho menos de lo que se conoce después de que comenzásemos a garabatear símbolos en tablillas de barro. La escritura nos hace partícipes del pasado y nos permite escribir el presente y el futuro. La de la escritura fue una revolución apacible (Vallejo, 2020:112) e inevitable una vez que el ser humano aprendió a hablar:

> Hace seis mil años, aparecieron los primeros signos escritos en Mesopotamia, pero los orígenes de esta invención están envueltos en el silencio y el misterio. Tiempo después, y de forma independiente, la escritura nació también en Egipto, la India y China. El arte de escribir tuvo, según las teorías más recientes, un origen práctico: las listas de propiedades. Estas hipótesis afirman que nuestros antepasados aprendieron el cálculo antes que las letras. La escritura vino a resolver un problema de propietarios ricos y administradores palaciegos, que necesitaban hacer anotaciones porque les resultaba difícil llevar la contabilidad de forma oral. El momento de transcribir leyendas y relatos llegaría después. Somos seres económicos y simbólicos. Empezamos escribiendo inventarios, y después invenciones (primero las cuentas; a continuación los cuentos). (Vallejo, 2020:113)

La lectura y la escritura son, pues, dos artificios de nuestra especie. Parece posible concebir una forma de evolución humana en la que expresión y comprensión oral se desarrollan sin expresión y comprensión escrita y, sin embargo, es difícil concebir lo contrario.

La escritura tiene sus propias características. No se adquiere, se aprende. Es permanente, demanda tiempo, permite la distancia entre el escritor y el lector y, por lo general, tiene mayor densidad léxica y gramatical que la producción oral (Weigle, 2002:15–16). A su manera, es una herramienta de expresión más refinada que la producción oral y, tal vez por eso, el medio que escogemos para expresar de forma exacta nuestras reflexiones:

> Poco después llegó un e-mail de Norton. Le pareció extraño que Norton le escribiera y no lo llamara por teléfono. A poco de leer la carta comprendió que Norton necesitaba expresar de la manera más ajustada posible sus pensamientos y que por esa razón había decidido escribirle. (Bolaño, 2004:64–65)

Hoy sabemos que nuestra capacidad de escribir no es estática, sino que evoluciona con el tiempo, en función de nuestro entorno y de los elementos tecnológicos a nuestro alcance (Bazerman *et al*. 2017). Quienes tras aprender a escribir a mano aprendimos a usar una máquina mecanográfica aún recordamos los cambios que nos permitió el posterior desarrollo de los editores de texto. Gracias a ellos la escritura dejó en parte de ser lineal y pudimos editar la primera línea de un texto sin necesidad de reescribir todas las demás. Está por ver qué supondrán los nuevos elementos tecnológicos para nuestra habilidad de escritura en particular y para la evaluación de lenguas en general. Los indicios sugieren que los correctores ortográficos basados en inteligencia artificial, por ejemplo, fomentan nuestra implicación emocional y cognitiva, y son de gran ayuda para quienes tienen problemas de escritura (Edmett *et al*. 2023:16). Es cuestión de tiempo que los *chat-bots* se implementen en los exámenes de producción escrita para generar interacción con el candidato en una primera fase, y para calificar la producción en otra posterior.

En las secciones que siguen analizaremos algunos de los aspectos fundamentales que se ha de tener en cuenta en el diseño de tareas de producción escrita. Al contrario de lo que ocurre en comprensión auditiva, comprensión lectora y producción oral, en el diseño de tareas para la evaluación de producción escrita no existe un catálogo consensuado de formatos de entre los que elegir. Existen distintos elementos (Weigle, 2002:90–107) que, combinados según las necesidades específicas de nuestro constructo, pueden dar lugar a multitud de tareas distintas. Lo que sigue es una descripción de los elementos que consideramos más relevantes en el diseño de tareas de producción escrita. En esta descripción no incluimos

las escalas de corrección de tareas, muy relevantes, ya que les dedicamos el capítulo 4 al completo.

3.4.4.1. Instrucciones

Las instrucciones deben ser una especie de resumen de todos los elementos que el candidato ha de tener en cuenta. Al igual que ocurre en el resto de destrezas, las instrucciones han de ser claras y deben dar indicaciones explícitas de aquello que se espera del candidato. Estas han de estar redactadas de forma que sean comprensibles, por lo que suele ser habitual redactarlas en un nivel de dominio inferior al del examen. Así pues, en un examen de dominio C1 las instrucciones se escribirían en un nivel B2; en un examen B2 se escribirían en nivel B1, etc. En el caso de pruebas multinivel, las instrucciones deben adaptarse al nivel más bajo que evalúe la prueba.

Algunos de los elementos que deben contener las instrucciones son 1) el tiempo de que los candidatos disponen para la tarea, 2) el número máximo de palabras que pueden escribir, 3) si tienen la opción o no de elegir entre varias tareas y, como decíamos en el párrafo anterior, 4) los diferentes puntos que se han de tratar en la respuesta.

Este último aspecto es de particular relevancia, dado que el cumplimiento de la tarea suele ser un criterio de corrección y, en ocasiones, se redacta de forma muy ambigua. Si, por ejemplo, proponemos «escriba un ensayo de X palabras sobre el tema Y enumerando los pros y los contras», el candidato podrá ser tan sucinto como quiera. Por ello, es recomendable dar indicaciones más precisas: «Escriba un ensayo de 300–350 palabras sobre el uso de las tecnologías aplicadas al aprendizaje de idiomas, enumerando al menos tres pros y tres contras, explicando su experiencia personal en este ámbito e indicando cómo cree usted que evolucionará la relación entre aprendizaje de idiomas y tecnología en el futuro». Mediante instrucciones tan detalladas como estas no solo explicitaremos todos los elementos que se tendrán en cuenta en el cumplimiento de la tarea, sino que, además, podemos favorecer la producción de cierto tipo de lenguaje. Por ejemplo, al pedirle al candidato que escriba sobre su experiencia personal, probablemente este usará tiempos verbales de pasado y presente, mientras que al hacer predicciones tendrá que usar tiempos futuros y tal vez el modo subjuntivo, si es que este existe en la lengua evaluada.

Por último, las instrucciones han de contextualizar la tarea de forma adecuada. La contextualización influye en la motivación del candidato y en cómo afronta la tarea. No es recomendable plantear una tarea como una mera operación lingüística. Lo ideal es revestirla de cierto parecido con cualquier actividad de la vida real en la que el candidato pueda verse reflejado, como hemos hecho en las tareas de las secciones 3.4.2 y 3.4.3.

3.4.4.2. Ámbito, tema y género

Según el MCER, los ámbitos son «los sectores amplios de la vida en los que actúan los agentes sociales» (Consejo de Europa, 2002:10). A su vez, «[l]as actividades de la lengua se encuentran contextualizadas dentro de ámbitos [...] que se pueden clasificar de forma general en cuatro: el *ámbito público*, el *ámbito personal*, el *ámbito educativo* y el *ámbito profesional*» (*ibid*.,15). Por lo tanto, será importante determinar en cuál de estos ámbitos se contextualizan las tareas que requerimos de nuestros candidatos, porque de ello dependerá el tema sobre el que tengan que escribir.

Existen exámenes orientados a ámbitos concretos y otros de carácter más general. Por ejemplo, un hipotético examen de dominio de lengua para controladores aéreos estaría centrado en el ámbito profesional. Exámenes, como DELE (Instituto Cervantes, 2014a; 2014b; 2019; 2022), TOEFL (Swain *et al*., 2009), etc., están orientados a una audiencia más general y, por tanto, podrán combinar el ámbito público y personal. En estos casos, los temas sugeridos para la producción escrita no pueden ser excesivamente específicos. Por ejemplo, en un examen de corte generalista no deberíamos incluir una tarea que esté orientada a un ámbito profesional o académico específico (aeronáutica, bioquímica, musicología, etc.), ya que podríamos estar perjudicando a determinados candidatos que, dadas las características del examen, asumen que temas tan específicos no se tratarán en la prueba.

En ocasiones, los propios nombres de los exámenes orientan sobre el dominio en el que se enmarcan, como ocurre, por ejemplo, con *IELTS Academic* o *Aptis ESOL for teachers*. La información sobre el ámbito y los temas que pueden formar parte del examen han de constar en las especificaciones y ser accesibles para los candidatos. De hecho, es incluso recomendable que en la versión de las especificaciones a las que pueden acceder los candidatos se incluya una lista de temas en los que podrían enmarcarse sus tareas de producción escrita. El MCER ofrece en su cuadro 5 (Consejo de Europa, 2001:48–49) y en su sección 4.2 (*ibid*., 51–53) una serie de categorías descriptivas y temas de comunicación que pueden servir de orientación a la hora de plantear esta lista (vida diaria, tiempo libre, viajes, salud, educación, compras, comida y bebida, servicios, etc.).

Según Weigle (2002:96–97), el género de la tarea de producción imbrica forma y función. Por forma entendemos el tipo de texto que el candidato habrá de producir (un correo electrónico, la reseña de un libro, un ensayo, etc.). La función, por otro lado, está relacionada con las microfunciones (ofrecer y buscar información factual, expresar actitudes, persuadir, sugerir, saludar, etc.), macrofunciones (describir, narrar, comentar, exponer, explicar, demostrar, argumentar, etc.) y esquemas de interacción que el candidato ha de usar (*vid*. Consejo de Europa, 2001:125–127).

Cuanto más claros estén estos elementos en las especificaciones y en las instrucciones, tanto mejor para el redactor de tareas, para el candidato y para el corrector. Por un lado, el redactor tendrá una lista acotada de temas y tipos de prueba sobre la que podrá dar instrucciones claras al candidato. Por su parte, el candidato sabrá a qué atenerse durante la preparación de la prueba. Por último, los calificadores se beneficiarán de las instrucciones claras dadas por los redactores, ya que los descritos por estos, y no otros, serán los aspectos que tendrán que buscar en la producción del candidato para evaluar el cumplimento de la tarea.

La tarea de la figura 16 de más abajo, por ejemplo, pertenece al ámbito personal. Como vemos, en las instrucciones se indica la duración de la prueba, el número máximo de palabras que el candidato puede usar y se le dan pautas concretas sobre las microfunciones, macrofunciones y esquemas de interacción que ha de utilizar.

3.4.4.3. Estímulo previo

Usar un estímulo previo no es estrictamente necesario, pero puede enriquecer la tarea de varias maneras.

Cuando facilitamos a los candidatos un estímulo previo, como por ejemplo en la tarea de la figura 16, nos aseguramos de que todos los candidatos parten exactamente del mismo contexto. Mediante esta contextualización propiciamos que el candidato se concentre en su producción lingüística, ya que no tendrá que imaginar una situación concreta (*vid*. Weigle, 2002:94–96) y contribuimos además a que active determinados esquemas mentales. Cuando facilitamos estímulos previos (como un texto, una imagen, un vídeo o un audio) las interferencias en la producción lingüística del candidato, no obstante, pueden provenir de su capacidad de interpretar dicho estímulo. Así pues, si se opta por usar estímulos previos, estos habrán de ser fáciles de interpretar.

Esta última reflexión sobre posibles interferencias entronca con las consideraciones que realizamos en la sección 3.4.2.8, en la que, al hablar sobre las destrezas integradas, nos planteábamos cuáles eran los procesos cognitivos que los candidatos utilizan (Plakans, 2008), en qué forma estos activan determinadas estrategias (Swain *et al*. 2009), y si la puntuación de un candidato en una tarea integrada refleja realmente su competencia en todas las destrezas que intervienen (Plakans, 2012:254). A pesar de las incógnitas, en nuestra opinión el estímulo previo reviste a estas tareas de autenticidad. En el día a día raramente nos lanzamos a escribir de forma descontextualizada o sin un estímulo previo. Cuando redactamos un correo electrónico lo hacemos porque queremos responder a otro anterior o porque tenemos una información que deseamos compartir. Cuando escribimos un mensaje de texto lo hacemos respondiendo a otro anterior o porque tenemos una intención comunicativa muy concreta. Incluso cuando hacemos una publicación en redes sociales utilizamos el texto para describir la imagen o el vídeo que compartimos.

Tarea de producción escrita

Recientemente ha comprado un teléfono móvil a través de internet. Al recibir su pedido observa que el teléfono no es nuevo sino reacondicionado y, tras escribir al vendedor, recibe usted el correo electrónico de más abajo.
Escriba una contestación en **250-300** palabras a este correo electrónico tratando todos los temas indicados.

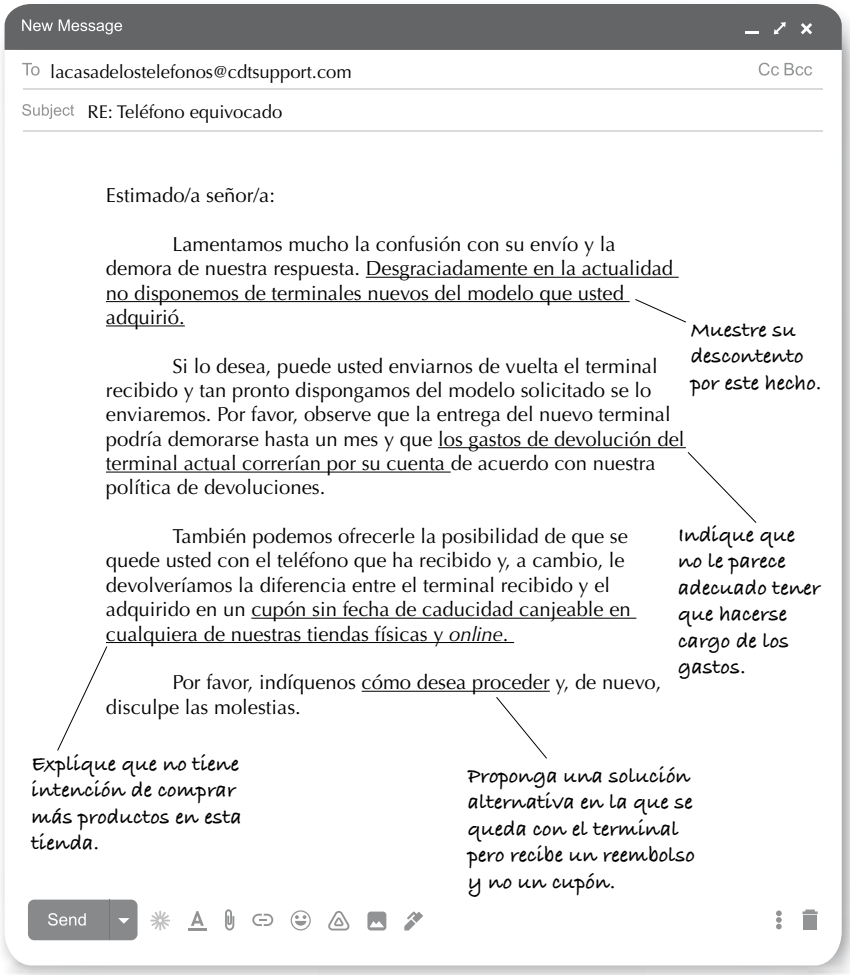


Figura 16. Tarea de producción escrita

3.4.5. Producción e interacción orales

Además de evanescente, la palabra hablada es compleja en tanto que requiere de múltiples procesos neuronales y biomecánicos (hasta 225 activaciones musculares por segundo según MacNeilage (2008:4)), y tan poderosa que, según la *Biblia*, la simple pronunciación de una frase dio luz a la humanidad. *Fiat lux*.

Las primeras manifestaciones orales de nuestra especie probablemente estuvieron ligadas a movimientos mandibulares cíclicos que el ser humano habría reciclado de su capacidad para masticar comida y que, más tarde, evolucionaron hacia una especie de clics sin apoyo vocálico similares a los que hoy día se registran en la familia de lenguas joisanas (o khoisánidas) de Bostwana (Poulos, 2022). En algún momento, estos sonidos habrían comenzado a acompañarse de vocales para originar las primeras protosílabas. Lo más probable es que estos movimientos mandibulares acompañados por vocales, también producidos de forma involuntaria por los bebés cuando succionan leche del seno materno, fuesen identificados por alguna madre de hace 70 000 años como una llamada de atención de su bebé hacia ella. La madre quiso ver un rasgo inédito de inteligencia en su recién nacido y el recién nacido, sencillamente, vio que repitiendo aquellos movimientos y sonidos era capaz de atraer la atención de su madre (MacNeilage: 2008:293). Esto podría explicar por qué las palabras *mamá* o *papá* son tan parecidas en lenguas que no tienen nada que ver entre sí (Jakobson, 1962:538–545). El ser humano parece estar predispuesto a dotar de significado a los sonidos.

Desde un punto de vista evolutivo observamos que nuestros ancestros no solo utilizaron esta capacidad de producir sonidos y vincularlos con significados sino que, además, a lo largo de millones de años aprovecharon diferentes elementos preexistentes en el cuerpo humano (cerebro, pulmones, lengua, labios, etc.) y, combinándolos, crearon una herramienta que les ayudó a sobrevivir y a convertirse en la especie dominante del planeta. El lenguaje es, probablemente, el hecho diferenciador que permitió a nuestra especie salir de la sabana.

> ¿Cuál fue el secreto del éxito del Sapiens? ¿Cómo logramos asentarnos tan rápidamente en tantos y tan distintos hábitats? ¿Cómo logramos hacer que otras especies humanas se extinguieran? […] Lo más probable es que la respuesta esté en aquello que hace este debate posible: el Homo sapiens conquistó el mundo principalmente gracias a su lenguaje único. (Harari, 2015:26)

Desde el punto de vista de la ontogenia, hoy sabemos que los biomarcadores presentes en la voz humana sirven, entre otras cosas, para detectar enfermedades mentales (Edmett *et al*., 2023:57). Como también mencionábamos en el capítulo 1, Tena *et al*. (2021) han demostrado que el análisis de la forma en que pronunciamos las vocales puede conducir a una detección temprana de la esclerosis lateral

amiotrófica, la enfermedad que postró en una silla de ruedas al físico Stephen Hawkins.

La producción oral es a al mismo tiempo el haz de la comprensión auditiva (y por tanto, comparte con esta mucho de lo descrito en la sección 3.4.2) y una destreza cuya evaluación requiere de consideraciones específicas: 1) el análisis síncrono o asíncrono del desempeño del candidato, que se realizará mediante 2) un tipo de tarea concreta, y que 3) se medirá con las escalas de corrección que aplican 4) unos evaluadores debidamente entrenados (Ginther, 2013).

El análisis síncrono de la producción de los candidatos es el más habitual, probablemente porque se ahorra tiempo, dado que se evalúa a los candidatos al tiempo que estos producen. En caso de que haya un solo evaluador (puede haber más), este puede tener que conducir la prueba al tiempo que evalúa al candidato. Si se requiere que el evaluador se centre exclusivamente en conducir la producción oral del candidato, o si se desea tomar medidas de precaución por si fuera necesario revisar la prueba en el futuro, esta puede ser grabada. Las pruebas de producción oral también son grabadas cuando el tipo de tarea (del que hablaremos a continuación) lo demanda, por ejemplo, en el caso de pruebas con tareas semidirectas realizadas a través de soportes digitales (ordenadores, tabletas, teléfonos móviles, etc.). Cuando la prueba es grabada para una evaluación posterior, el evaluador tiene la posibilidad de escucharla en repetidas ocasiones sin las limitaciones cognitivas que supone desempeñar dos papeles (el de conductor y evaluador) al mismo tiempo.

Las tareas de producción oral pueden ser directas, indirectas y semidirectas (Clark, 1979:36). Los métodos de evaluación oral directos son aquellos que requieren que el candidato realice un intercambio verbal, cara a cara, con uno o más interlocutores. Este tipo de evaluación es, sin duda, la más frecuente en los exámenes de dominio, probablemente por su similitud con la forma en que la palabra hablada se usa en la vida real. Los métodos indirectos son aquellos que, aunque pueda parecer un contrasentido, no requieren que el candidato hable. En esta categoría entraría, por ejemplo, una tarea en la que el candidato tiene que discernir qué palabra de una lista se pronuncia diferente al resto. Finalmente, los métodos semidirectos son aquellos en los que el candidato tiene que hablar a partir de estímulos no humanos tales como grabaciones o imágenes. Hoy en día los métodos indirectos están en desuso en el ámbito de los exámenes de dominio y la barrera entre sistemas directos y semidirectos está desdibujada por el uso de inteligencia artificial, como veremos en la sección 3.4.5.3.

Fulcher (2003:57) considera la forma de intercambio verbal (directa, indirecta o semidirecta) una característica más de las tareas, no la principal, y la incluye en una taxonomía que permite describir cualquier tarea más allá del tipo de interacción que genera. El marco descriptivo que propone Fulcher (*ibid.*) es el siguiente:

1. Orientación de la tarea
    - Abierta
    - Guiada
    - Cerrada
2. Tipo de interacción
    - Sin interacción
    - Con interacción
3. Objetivo
    - Ninguno
    - Convergente
    - Divergente
4. Familiaridad y estatus del interlocutor
    - Sin interlocutor
    - Estatus superior
    - Estatus inferior
    - Mismo estatus
    - Grado de familiaridad con el interlocutor
5. Temas
6. Situación

Tabla 11. Marco descriptivo de tareas de producción oral

En las siguientes secciones analizaremos algunos tipos de tareas partiendo de la clasificación de Clark (1979:36) y recuperaremos algunos de los ejemplos propuestos por Fulcher (2003:50–87). De las escalas y los evaluadores nos ocuparemos en el capítulo 4.

3.4.5.1. Tareas directas

Como ya hemos comentado anteriormente, las tareas directas son aquellas que requieren que el candidato realice un intercambio verbal, cara a cara, con uno o más interlocutores (Clark, 1979:36). Las que se encuentran con más frecuencia son las basadas en preguntas personales.

Las tareas basadas en preguntas personales suelen ubicarse al principio de la prueba de producción oral, ya que favorecen que el candidato se acomode a la misma. Dado que suelen estar asociadas con temas familiares para los examinandos, son una buena forma de romper el hielo y mitigar el nerviosismo inicial durante las pruebas. Pensemos que para evaluar la producción oral de nuestros candidatos les pedimos que hablen en una lengua que no dominan, con una o varias personas a las que jamás antes han visto, y de cuyo juicio depende su calificación. Una situación que, sin duda, a la mayoría nos causaría nerviosismo. La variedad de preguntas que se pueden realizar es infinita, si bien la experiencia nos

dice que la mejor forma de romper el hielo y ayudar a los candidatos a mitigar su nerviosismo es la de centrarse en el ámbito personal, ocupacional o educativo, y no tanto en el público (*vid*. Consejo de Europa, 2001:14–15).

3.4.5.2. Tareas indirectas

Como hemos comentado más arriba, este tipo de tareas está actualmente en desuso en la evaluación de dominio de lenguas. Se utilizan fundamentalmente en niveles muy bajos de dominio o con fines muy concretos que suelen circunscribirse a la evaluación de la ortoepía, es decir, de la capacidad de pronunciar correctamente los sonidos de una lengua.

Estas tareas pueden tener la forma una lista de palabras de entre las que se pide al candidato que señale aquella que se pronuncia con un sonido distinto al de las otras. Pensemos, por ejemplo, en una tarea en la que se le dan al candidato las palabras «gala, «agosto», «gente», «glaciar» y «gueto» y se le pide que identifique en cuál de ellas la letra «g» se pronuncia de forma distinta. Otra de las formas que pueden adoptar este tipo de tareas es aquella en la que el evaluador pide al candidato que repita distintas oraciones de longitud variable para evaluar no solo la pronunciación aislada de sonidos sino también la entonación (útil para lenguas tonales como el chino), el ritmo y la fluidez.

3.4.5.3. Tareas semidirectas

Se consideran tareas semidirectas aquellas en las que el candidato tiene que hablar a partir de estímulos no humanos tales como grabaciones o imágenes (Clark, 1979:36). En este grupo encontramos, por ejemplo, las tareas en las que se ha de realizar una descripción, una comparación de imágenes, una narración o establecer un debate con otros candidatos. Todas ellas suelen darse en un formato de entrevista, si bien la participación del examinador es mínima y suele estar limitada a la conducción de la prueba, de ahí que las consideremos semidirectas y no directas.

Las tareas basadas en la descripción de imágenes generalmente están orientadas a que el candidato produzca un monólogo, y son muy frecuentes en exámenes de niveles bajos o intermedios, ya que favorecen el uso de los conectores y tiempos verbales más sencillos. En el caso de que haya varios candidatos en la misma prueba, para evitar que los que no participen en el monólogo se sientan excluidos o pierdan la concentración, suele ser habitual hacer una pregunta sobre la intervención del compañero al candidato que ha permanecido silente durante la comparación de imágenes. Lo habitual es pedir al candidato A que describa lo que está ocurriendo en una única imagen (dibujo o fotografía), que habrá de tener suficiente contenido para generar la producción del candidato. Por ejemplo, la imagen de un cielo azul y despejado puede ser muy atractiva, pero contiene

poca información que se pueda describir. Es necesario, pues, que estas imágenes contengan objetos, personajes que realicen acciones, etc. Para aumentar las posibilidades de descripción del candidato se puede usar una variante de esta tarea en la que se le pide que describa y compare dos imágenes sobre una misma temática con perspectivas distintas, como en la figura 17. Después de que el candidato A haya concluido se le puede hacer una pregunta al candidato B sobre lo que el primero ha dicho y viceversa. Todas estas interacciones han de estar pautadas y guionizadas en la medida de lo posible para evitar que el parafraseo de un evaluador pueda dar ventajas con respecto al parafraseo de otro.

Tarea de producción oral

Estas dos fotografías representan diferentes tipos de viviendas. Describa y compare durante **2 minutos** ambas fotografías, indicando los **pros y contras** de vivir en uno u otro sitio. Para finalizar, **indique** en cuál de estos tipos de vivienda preferiría usted vivir.

Figura 17. Tarea de producción oral (I)

Otra variante es la que incluye diferentes imágenes que, en su conjunto, describen una historia que se pide al candidato que narre. En niveles más elevados se le puede solicitar la descripción de una gráfica, entre otras tareas.

Son bastante populares las tareas que proponen un debate a dos o más candidatos. Estas tareas nos permiten medir habilidades como la capacidad de gestionar los turnos de palabra o de negociar para llegar a un acuerdo. Dado que conforme avanza el debate los candidatos intervienen respondiendo a lo dicho por su interlocutor, estas tareas pueden considerarse como un híbrido de tareas directas y semidirectas.

Tarea de producción oral

Este año ha pensado pasar las vacaciones entre amigos y ha quedado usted con uno de ellos para decidir cuál de las diferentes opciones que tienen es la mejor. Charle con su compañero/a durante **3 minutos** sobre las opciones de más abajo, y decida cuál es la mejor para sus próximas vacaciones juntos. Recuerde que al final de la tarea tendrá que llegar a un acuerdo con su interlocutor/a.

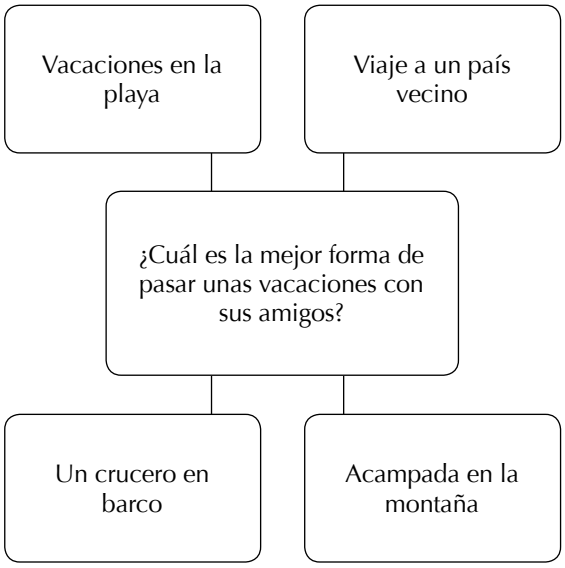


Figura 18. Tarea de producción oral (II)

En sentido estricto, según la definición de Clark (1979:36), en las tareas semidirectas el candidato interactúa a partir de estímulos no humanos. Creemos que es interesante en este punto reflexionar sobre el lugar que ocupará la inteligencia

artificial a medio y largo plazo en la evaluación de destrezas productivas (tanto orales como escritas). Hoy día, la voz e incluso el rostro de múltiples inteligencias artificiales (que no son un estímulo humano) se ha colado en nuestra cotidianidad. Interactuamos a diario con ellas, de viva voz, para pedirles que programen una ruta en el sistema de navegación, que busquen algo por nosotros en internet o incluso para pedirles que pongan música cuando llegamos a casa después de un duro día de trabajo. Existen indicios de que la inteligencia artificial aplicada a la enseñanza puede ayudar a conseguir una pronunciación más natural (Liu y Hung, 2016) y fluida (Shivakumar *et al.* 2019), a retener vocabulario de forma más efectiva (Kazu y Kuvvetli, 2023), y que puede propiciar un contexto de interacción significativa que hace más interesante el aprendizaje (Dizon y Tang, 2020). En nuestra opinión, al igual que ocurrirá con las pruebas de producción escrita, es cuestión de tiempo que se incluyan *chat-bots* en las pruebas de producción oral implementadas con ordenador (o dispositivos digitales similares) que, en primera instancia, ayudarán a fomentar una interacción significativa y que, con posterioridad, permitirán la corrección semiautomática de la producción del candidato.

3.4.6. Otros modelos

Con esta sección cerramos el repaso a los distintos componentes de los exámenes de dominio de lengua. La estructura aquí analizada, que se basa en la distinción tradicional de destrezas iniciada por Lado (1961), es a la que ha conducido el desarrollo de la evaluación objetiva de lenguas desde mediados del siglo XX, y parece seguir siendo el punto de partida de la mayoría de pruebas existentes en el momento en que redactamos este libro.

Existen, como es obvio, otros modelos de evaluación de lenguas. Por ejemplo, están aquellos modelos que incorporan el uso de la lengua (esto es, gramática y vocabulario) como un componente individual. Estos modelos, muy populares durante décadas, hoy en día se encuentran relegados a casos particulares como, por ejemplo, la *suite* de exámenes de Cambridge University Press and Assessment, en los que, aunque la evaluación del uso de la lengua se presenta junto a la comprensión lectora, constituye una parte distinta de esta. De hecho, la tendencia actual parece apuntar en sentido opuesto: no se busca diseñar exámenes de competencias aisladas sino, más bien, diseñar exámenes que integren destrezas, competencias o modos de comunicación, como ocurre en la evaluación basada en escenarios (Purpura, 2021).

Son particularmente interesantes los modelos orientados al plurilingüismo que, como North y Piccardo (2023:10) nos recuerdan, ya han demostrado ser útiles para evaluar cómo el hablante usa lenguas conocidas en relación con otras que no conoce (Rehbein *et al.*; 2012; Galante, 2018; Jentges *et al.* 2023) e incluso para analizar el papel de las lenguas de herencia en el contexto de la evaluación

plurilingüe (Escamilla *et al*. 2013; González-Davies, 2020; Piccardo *et al*. 2022). Tal vez el más conocido de todos estos modelos plurilingües sea el de las pruebas que se realizan en la enseñanza profesional secundaria austríaca[2] (Atzlesberger *et al*., 2015; Steinhuber, 2022).

Con el plurilingüismo de fondo, modelos basados en los escenarios como el del certificado CLES[3] del Ministerio de Educación francés tienen como objetivo involucrar a los candidatos en la resolución de un problema dentro de los parámetros contextuales de una narrativa que no solo les resulte atractiva sino que, además, les facilite una experiencia educativa valiosa por sí misma (*vid*. Purpura, 2021:A74). Los escenarios pueden ser una magnífica fuente de inspiración para nuestros exámenes en tanto que aglutinan tareas orientadas a la acción que empujan a los aprendientes (o candidatos) a colaborar para crear *artefactos* de diversa índole (textos orales, escritos o productos multimedia vinculados con códigos semióticos) (Piccardo y North, 2019:272), a través de los cuales se puede medir el dominio lingüístico de uno o varios idiomas.

Tanto los modelos plurilingües como los basados en escenarios suponen retos adicionales en el diseño de tareas. Por ejemplo, es necesario definir cuántas lenguas se pueden usar en la prueba, cuáles se consideran primera o segunda lengua, etc. Además, no todos los ejemplos mencionados más arriba son practicables en contextos de evaluación estandarizada. Con todo, y a pesar de los retos, la orientación plurilingüe de cualquier examen de dominio tiene particular interés en un continente como el europeo, en el que no existe una única lengua común, al contrario de lo que ocurre en la mayor parte de Latinoamérica.

3.5. El ciclo de diseño de un examen

Un buen examen de dominio se comienza a crear mucho antes de redactar el primer ítem. Es importante saber que se dispone de los recursos (personales y materiales) necesarios para el diseño. Es crucial que se haya hecho una reflexión previa sobre el constructo de la prueba y que se hayan definido unas especificaciones (*vid*. 3.4.1) que estén disponibles con suficiente antelación para los redactores y los candidatos. Solo entonces será posible comenzar a diseñar un examen de dominio de lenguas.

Una de las mejores formas de establecer un flujo de trabajo efectivo es crear un procedimiento, una lista de tareas que se deban completar de forma secuencial y tener claro qué se ha de hacer en cada una de sus diferentes fases. Partiendo del ciclo propuesto por Green (2017:21–25), en esta sección, resumida en la figura 19

2. https://www.cebs.at/home/plurilingualism
3. https://www.certification-cles.fr

de más abajo, enumeraremos los hitos principales en el diseño de una prueba de dominio de lengua.

En una situación ideal, en la que se contase con tiempo suficiente y un número ilimitado de recursos, todos los hitos descritos deberían tener la misma relevancia en el diseño de nuestras pruebas. No obstante, la realidad es obstinada, los recursos suelen ser limitados y, con frecuencia, los redactores de exámenes de dominio tenemos que armonizar aspectos tan distantes entre sí como la lingüística y la psicometría al tiempo que lidiamos con presiones institucionales, sociales y económicas (Spolsky, 1995:4). Por estos motivos, si bien hemos de apuntar siempre a la excelencia y a observar cuidadosamente cada uno de los pasos descritos, también hemos de adaptarnos a la realidad de nuestro contexto y nuestros recursos.

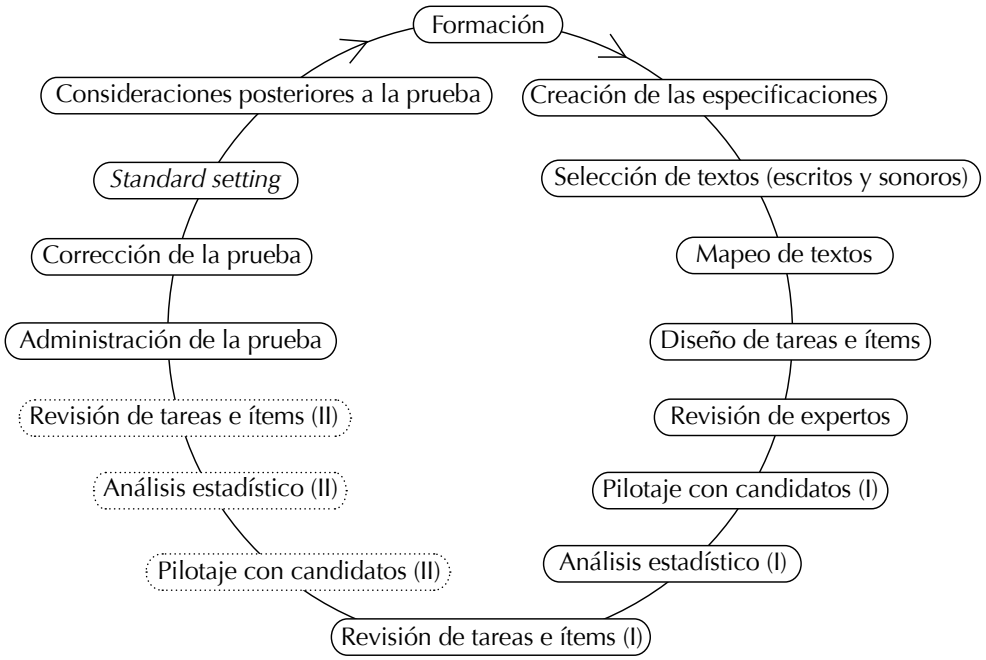


Figura 19. Ciclo de diseño de un examen

3.5.1. Selección de textos (escritos y sonoros)

El nivel de dificultad de nuestra prueba es una suma de la dificultad del texto escogido y de la de los ítems y tareas confeccionadas.

Los principales parámetros que debemos tener en cuenta a la hora de escoger textos son su extensión y nivel. En nuestras especificaciones debería estar

identificada la extensión máxima y mínima de los textos que se eligen, así como su número, si las tareas constarán de uno o varios textos, etc. Es difícil encontrar textos que se adapten a un único nivel de dominio. Con todo, el nivel general del texto escogido debe corresponder lo máximo posible con aquel que estamos evaluando. Para un nivel B1, escogeremos textos de nivel B1, para un examen de nivel C1 escogeremos textos de nivel C1, etc., aunque hemos de asumir que cualquier texto de este tipo compartirá características de otros niveles de dominio.

Los textos seleccionados, además, han de ser relevantes temáticamente. Cuanto más factible sea que un candidato encuentre un texto del tipo elegido en la vida real, mayor validez de apariencia (*face validity*) (Davies *et al.*, 1999:59) tendrá para este la prueba. Internet es una fuente inagotable de todo tipo de textos. En el caso de comprensión auditiva, por ejemplo, se pueden encontrar multitud de *podcasts* de diferentes temáticas. También existen páginas especializadas en diferentes materias (redes sociales, alquiler de viviendas, comercio electrónico, etc.) que generan su propio contenido y que podremos utilizar para cualquier tipo de tarea. Por ejemplo, podemos recopilar varias opiniones sobre distintos alojamientos obtenidos en una *web* de alquiler de viviendas para generar un ejercicio de emparejamiento de comprensión lectora. Podemos separar el audio del vídeo de una grabación alojada en una *web* que aglutine conferencias para crear un ejercicio de toma de notas de comprensión auditiva. Las posibilidades son casi infinitas.

En la medida de lo posible, los textos elegidos han de ser auténticos. Dado que no existe una definición unificada de lo que es la autenticidad en la evaluación de lenguas (Hasrol *et al.*, 2022), en las secciones que siguen entenderemos autenticidad como el grado de correspondencia entre el tipo de lengua que se usa en un contexto determinado de la vida real y la que usamos en nuestras pruebas (*vid*. Lewkowicz, 1996; 2000 y Green, 2017:37–38). Desde esta perspectiva, la autenticidad no es dicotómica, no es una cuestión de «todo o nada», sino que se da en mayor o menor grado en un texto. En un extremo de esta gradación encontraríamos exámenes compuestos por textos sin modificar, extraídos directamente de su contexto original, y en el otro extremo encontraríamos exámenes compuestos por textos creados *ex profeso* para la prueba.

Un texto sin modificar extraído directamente de contextos reales puede no ser necesariamente la mejor opción para nuestros exámenes. Dado que los autores de estos textos no los concibieron para la evaluación de lenguas, pueden contener una densidad léxica o gramatical desigual y ser más extensos de lo que es necesario para nuestras pruebas. Es muy difícil que un texto de este tipo se adapte a un único nivel de dominio. Casi cualquier texto que escojamos para un nivel concreto contendrá rasgos de niveles de dominio adyacentes. Tampoco se ha de caer en la falsa creencia de que cualquier texto redactado por un hablante nativo es apto para una prueba de dominio en tanto que es auténtico. De hecho, la idea del hablante nativo como referente de corrección está en desuso, como lo

demuestra el hecho de que en el MCERVC (Consejo de Europa, 2020) se eliminaran las trece referencias a hablantes nativos que se habían incluido en el MCER (Consejo de Europa, 2001).

Desde nuestro punto de vista es adecuado generar textos *ex profeso* para un examen (*vid*. 3.4.2). En ocasiones, ante la dificultad de encontrar textos de la temática, extensión y el nivel correcto para una prueba, la solución más práctica es generar nuestras propias grabaciones de audio o nuestros propios textos de comprensión lectora. Aunque pueda sonar contradictorio, generar nuestros propios textos puede ayudarnos a controlar mejor la autenticidad de una prueba, ya que, como expertos, por lo general, somos capaces de reproducir situaciones de escucha, lectura, habla y escritura que se dan de manera natural en la vida real.

Quizás la mejor opción sea optar por una de mezcla diferentes textos: textos extraídos de contextos reales ligeramente adaptados para nuestras pruebas (mediante la simplificación, la sustitución o anotación de palabras que no correspondan al nivel que estamos evaluando, etc.).

3.5.2. Mapeo de textos

Mapear consiste en identificar de forma colegiada los elementos más relevantes de un texto escrito o sonoro. Esta técnica, que no se usa en el diseño de todos los exámenes de dominio, presenta una serie de ventajas que describimos a continuación.

Con frecuencia, cuando el redactor de una tarea comienza a dar los primeros pasos para el diseño de ítems lo hace en solitario. Tras escoger un texto lo lee o escucha hasta dar con aquellos segmentos de información que en su opinión son susceptibles de acabar convirtiéndose en un ítem. El problema de esta aproximación individual es que la información que es relevante para un redactor puede no serlo para otro. En un proceso de mapeo se comparan las ideas que varios expertos consideran relevantes y allí donde hay coincidencias se identifican los segmentos susceptibles de generar ítems.

Aunque Green (2017:57-84) habla de tres tipos diferentes de mapeos, en la práctica se suelen usar dos: el mapeo para ideas principales y secundarias por un lado y, por otro, el mapeo para detalles específicos.

En los mapeos orientados a encontrar detalles específicos buscaremos fechas, horas, lugares, nombres, precios, porcentajes, números (de teléfono), medidas, acrónimos, direcciones, páginas de internet, etc. Para ello, tres o más expertos leerán o escucharán el texto por separado y una sola vez. El objetivo de escuchar o leer el texto una sola vez es reproducir los procesos cognitivos que el candidato empleará durante la prueba. Tras esta lectura o escucha, los expertos han de revisar sus anotaciones para eliminar aquellos aspectos que puedan no ser detalles específicos. Hecho esto, la persona encargada de recopilar la información del

mapeo, alguien distinto de los expertos, recopilará la información realizando dos tareas principales. Por un lado, identificará qué ideas han sido anotadas por los expertos. Por otro, anotará en qué línea del texto o segundo del audio aparecen estas. Toda la información se volcará en una tabla y solo aquellos segmentos en los que coincidan dos o más expertos serán susceptibles de ser tenidos en cuenta para generar ítems. El hecho de que la ubicación de la información se recoja en una tabla ayudará a distribuir adecuadamente los ítems que se diseñen y a que estos no se solapen.

Detalle específico	Experto 1	Experto 2	Experto 3	Ubicación
Detalle 1	x		x	0′10″
Detalle 2	x			0′17″
Detalle 3	x	x	x	0′25″
Detalle 4			x	0′33″
Detalle 5	x	x		0′47″
Detalle 6	x	x		0′53″
Detalle 7			x	1′07″
etc.				

Tabla 12. Mapeo de textos

En el ejemplo descrito en la tabla 12 hemos sombreado todos los detalles encontrados en los que coinciden dos o más expertos. Una vez identificados, hemos de elegir aquellos que tengan suficiente separación entre sí, por ejemplo, los detalles 1 y 3. Dado que los detalles 5 y 6 están muy cerca el uno del otro, habrá que decidir cuál se escoge para la redacción del siguiente ítem. Tan importante es que dos ítems no estén muy juntos como lo es que no estén muy separados. Una separación excesiva puede poner nerviosos a los candidatos que busquen respuesta a un ítem que crean haber perdido pero que, en realidad, no existe.

Todo en el mapeo para ideas principales y secundarias es igual al mapeo de detalles específicos, salvo el contenido que los expertos han de rastrear. Antes de comenzar el mapeo es importante dejarles claro lo que buscan. Las ideas principales son aquellas en torno a las que se articula un texto, mientras que las secundarias son las que sustentan las principales sin ser detalles específicos. Recordemos que el objetivo de una tarea de mapeo es encontrar ideas para diseñar ítems, no definir cómo está estructurado conceptualmente el texto, por lo que durante las sesiones de mapeo no merece la pena dedicar más tiempo del necesario al debate sobre qué es principal o secundario. Por ejemplo, en el texto sobre la cumbia de la

figura 15, las ideas principales podrían ser «el origen de la cumbia» y «la evolución de la cumbia». Las ideas secundarias podrían ser «los elementos etnoculturales en la cumbia», «su evolución instrumental y técnica» o «las variedades de cumbia existentes». Como decíamos, aunque pueda haber dudas acerca de si algunas de estas ideas pertenecen a lo principal o a lo secundario, lo importante es encontrarlas y diferenciarlas de detalles específicos como «la cumbia nace en el valle del río Magdalena» o «el origen de la cumbia rebajada se atribuye al DJ Gabriel Duéñez». Encontradas las ideas principales, estas se trasladarán también a una tabla similar a la 12 para la toma de decisiones.

La experiencia nos dice que las anotaciones que no llegan a convertirse en ítems son útiles tanto en el mapeo de detalles específicos como en el de ideas principales y secundarias. Por ejemplo, los detalles específicos descartados pueden usarse como distractores para los ítems que se abran paso a la versión final. Además, los expertos suelen parafrasear las ideas principales y secundarias al anotarlas, así que sus reformulaciones pueden usarse para la redacción de las raíces de los ítems. También puede ser interesante consultar la opinión de los expertos sobre la dificultad o extensión de los textos.

3.5.3. Diseño de tareas e ítems

Sin duda esta es una fase crucial del diseño de nuestras pruebas. Toda la información relativa a la forma de nuestras tareas ha de estar disponible en las especificaciones para quienes diseñan las pruebas y ser accesible para los candidatos.

Para los redactores será fundamental que las especificaciones contengan indicaciones precisas sobre aspectos como el número total de tareas de cada componente, el tipo y número de textos que incluirán (un solo texto extenso, una mezcla de textos más cortos, etc.), los diferentes formatos que se pueden usar (respuesta múltiple, emparejamiento, etc.) y el número de ítems que han de contener. En suma, las especificaciones han de contener plantillas de tareas que permitan a los redactores hacer el constructo operativo (Bachman y Palmer, 1996:171–180).

En la sección 3.4 incluimos indicaciones precisas y ejemplos de los diferentes tipos de tareas que se pueden diseñar para cada uno de los componentes de las pruebas, por lo que no los repetiremos aquí.

3.5.4. Revisión de expertos

Una vez elegidos y mapeados los textos, y confeccionadas las tareas e ítems, es recomendable que un grupo de expertos revise el trabajo realizado hasta este momento. La idea es que cada uno de estos expertos reproduzca todos los procesos cognitivos que tendrá que llevar a cabo un candidato y que realice la prueba como si fuera uno de ellos.

Para esto facilitaremos a los expertos una copia completa del examen o de las tareas que deseamos que revisen, que no incluirá las respuestas correctas, y les pediremos que realicen la prueba en el mismo tiempo en que tendría que realizarla un candidato. Al terminar, además de las respuestas que los expertos marquen como correctas, les pediremos que identifiquen cualquier inconsistencia que hayan detectado (errores de impresión, faltas de ortografía, preguntas que se pueden contestar sin necesidad de leer o escuchar el texto, etc.). Asumimos que, al tratarse de expertos, las respuestas que estos den a los distintos ítems deberían siempre coincidir con las que los redactores han identificado como correctas. Prestaremos particular atención a aquellos casos en los que no sea así, algo que puede ocurrir porque los expertos hayan cometiendo un error, porque el ítem acepte varias respuestas o porque la respuesta que creíamos correcta no lo sea realmente. Por regla general, si un ítem o tarea causa confusión o debate entre los expertos, también causará confusión entre los candidatos, algo que debemos evitar. Recibidas las respuestas de los expertos realizaremos las modificaciones oportunas.

Como hemos mencionado anteriormente, destinar más o menos recursos a esta fase de diseño dependerá del contexto en el que nos encontremos. Es recomendable que un examen sea revisado por un mínimo de tres expertos distintos y que uno de ellos sea nativo de la lengua que la prueba evalúa.

3.5.5. Pilotaje con candidatos

El objetivo de esta fase de diseño es recopilar información estadística sobre cómo funcionan nuestras tareas e ítems en un contexto parecido al real.

Para ello, pasaremos nuestra prueba a una muestra representativa de la población a la que esté dirigida. La muestra será representativa si tiene características similares a las de quienes más tarde serán los candidatos reales. Si, por ejemplo, la prueba está orientada a estudiantes universitarios, hemos de buscar candidatos para el pilotaje en instituciones universitarias; si está orientada al ámbito empresarial, hemos de buscar candidatos con el mismo contexto laboral que los que luego realizarán nuestras pruebas, etc. Además, la muestra de candidatos con la que contemos para el pilotaje habrá de contener candidatos con un nivel de dominio similar y adyacente al de nuestros potenciales candidatos. Si, por ejemplo, estamos diseñando un examen de dominio B2, lo ideal sería contar con candidatos de nivel B2 (el nivel para el que estamos diseñando la prueba) y con candidatos de nivel B1 y C1 (niveles adyacentes). En casos de exámenes multinivel, obviamente, cuantos más niveles represente la muestra, mejor.

No existe un consenso claro sobre cuán grande ha de ser el tamaño de la muestra de candidatos que usemos en el pilotaje. De nuevo, esto está sujeto a los recursos disponibles y al impacto de la misma. La literatura parece sugerir que a partir de los treinta y cinco candidatos se pueden obtener resultados aceptables y

que cincuenta candidatos bien escogidos es un número adecuado (Linacre, 1994; Wright y Tennant, 1994; Mustafa y Robillos, 2020) para determinados análisis.

Es recomendable analizar las respuestas de los candidatos del pilotaje al tiempo que se transcriben para su interpretación estadística. Los candidatos de los pilotajes suelen ser muy creativos, dado que para ellos aprobar o suspender la prueba no tiene el mismo impacto que para los candidatos finales, y esto les suele permitir tomarse ciertas licencias. Gracias a esta creatividad, a veces es posible encontrar variables correctas a las claves originalmente diseñadas.

3.5.6. Análisis estadístico

Un análisis estadístico sólido supone una fuente de validez muy importante para todos los agentes interesados en la prueba, desde los redactores a los candidatos. Gracias a los resultados obtenidos durante el pilotaje con candidatos conseguiremos una nueva perspectiva de nuestra prueba. Veremos dónde existen puntos de mejora y fruto de ello realizaremos las modificaciones pertinentes.

Para realizar este análisis es necesario contar con un psicometrista o, mejor aún, con un lingüista con conocimientos de psicometría. De nuevo, cuantos más recursos se destinen a este análisis, tanto mejor. Existen multitud de análisis que se pueden realizar, de los que hemos hablado ya en la sección 2.1. En la sección 3.3 también hemos visto cómo estos y otros análisis pueden ordenarse y utilizarse en la construcción de nuestro argumento de validez. La tabla 10 de la sección 3.3 se puede utilizar como guía para decidir qué análisis requieren nuestras pruebas. Existen múltiples manuales y artículos que permiten ir más allá de lo propuesto en esta tabla (Wright y Stone, 1978; McNamara, 1996; Eckes, 2009; Green, 2013; el tutorial de Linacre, 2024a y 2024b, etc.) si se cuenta con recursos suficientes.

3.5.7. *Standard setting*

Se conoce como *standard setting* al procedimiento por medio del cual se establece la calificación de corte (*cut-off score*) de una prueba. Esta calificación puede determinar el aprobado en una prueba uninivel o los umbrales que separan diferentes niveles de una prueba multinivel (*vid*. Davies *et al*., 1999:186). Para ello, *grosso modo*, se realizan consultas sistemáticas a un grupo de expertos sobre las puntuaciones de una prueba y sobre la vinculación de estas con los ítems y tareas de las que emanan. Las respuestas de los expertos son analizadas estadísticamente y fruto de estos análisis se establecen los mencionados puntos de corte.

El *standard setting* es un proceso complejo y costoso que ha atravesado varias fases de popularidad y complejidad, y que no está definido por una única metodología. En su exhaustiva revisión histórica y metodológica, Kaftandjieva (2004:11) habla de la existencia de más de cincuenta métodos distintos de *standard setting*, cada uno de los cuales puede conducir a calificaciones de corte

distintas. Esta arbitrariedad en los resultados (Kane, 1994:434), así como la complejidad técnica del proceso, hacen que su uso sea poco habitual, particularmente en pruebas de dominio con un número de candidatos o impacto limitados, en las que los puntos de corte pueden incluso estar preestablecidos por el currículo u otras normativas.

Realizar un proceso de *standard setting* para cada convocatoria no es practicable ni siquiera para las grandes casas examinadoras, que reconocen que solo se embarcan en procesos de esta envergadura cuando se producen cambios sustanciales en la estructura de sus pruebas. Quienes deseen profundizar en todo lo que supone el *standard setting* encontrarán la lectura de Kaftandjieva (2004) particularmente interesante.

3.5.8. Administración de la prueba

Todos los preparativos previos conducen a este momento crítico. La administración de la prueba puede tener lugar en una o varias sedes separadas por kilómetros. En el caso de que la administración tenga lugar en varias sedes es necesario asegurarse de que todos los candidatos concurren a la prueba en igualdad de condiciones. La ubicación de los puestos de los candidatos puede influir en el desempeño de estos si, por ejemplo, las condiciones de luz no son las adecuadas para la comprensión lectora, si están situados lejos de la fuente de sonido en una prueba de comprensión auditiva, etc. Es conveniente realizar varias comprobaciones de todas estas condiciones antes de la prueba, incluidas las relativas al sonido ya descritas en la sección 3.4.2.

El lugar de realización de las pruebas ha de haber sido comunicado a los candidatos con antelación, y el día de la prueba el acceso a las instalaciones debe estar señalizado, puesto que es posible que la mayoría de nuestros candidatos no conozcan el lugar. Para asegurar que las pruebas transcurren con normalidad es recomendable confeccionar una lista con aquellos aspectos que hay que tener en cuenta antes, durante e inmediatamente después de la prueba. Algunas de las consideraciones que deben tener estas guías para la administración de exámenes son:

Antes de la prueba

1. Comprobar qué personal se encargará de la vigilancia de la prueba.
2. Comprobar que se dispone de suficientes copias del examen.
3. En caso de que se trate de un examen por ordenador, realizar una comprobación del *hardware*, del *software* y de la conexión a internet en caso de que sea necesaria.
4. Comprobar que los dispositivos de reproducción de audio funcionan correctamente.

5. Preparar un listado de los candidatos para los vigilantes de la prueba, si es posible, con los datos de contacto de aquellos.
6. Disponer un reloj que sea visible para todos los candidatos durante la prueba.

El día de la prueba

7. Los encargados del desarrollo de las pruebas han de estar en las instalaciones con suficiente antelación y con la documentación y herramientas necesarias (listado de participantes, horarios, instrucciones, reloj visible para los candidatos, útiles de escritura, dispositivos electrónicos, auriculares, etc.).
8. Los encargados llevarán a cabo una segunda comprobación de *hardware*, *software* y conexión, así como de los sistemas de reproducción de audio.
9. Antes de que los candidatos accedan a las instalaciones se debe colocar en los distintos puestos el material necesario (identificadores, papel, auriculares, dispositivos electrónicos, etc.).
10. Los candidatos han de acceder a las instalaciones al menos quince minutos antes de la prueba para identificarse, dejar sus pertenencias en los lugares asignados y ocupar su puesto (se les debe avisar de esto en comunicaciones previas para que acudan puntuales y debidamente documentados). Es recomendable que mientras que una de las personas responsables los identifica y les da acceso a las instalaciones, otra les ayude a ubicarse en el lugar correcto.
11. Antes de comenzar, en la lista de candidatos se ha de anotar si hay alguna ausencia y el número total de candidatos presentes.
12. Una vez los candidatos están ubicados en sus puestos se les dan algunas instrucciones generales.
    - Se les recuerda que han de tener todos los dispositivos digitales propios apagados para evitar que llamadas, alarmas o notificaciones interfieran en el examen y, particularmente, en el componente de comprensión auditiva.
    - Se les recuerda la mecánica de la prueba: sus partes, duración estimada, cómo han de responder a las preguntas y si han de hacerlo con bolígrafo o con lápiz, si al marcharse han de entregar el examen completo, si han de dejarlo en su puesto para que los vigilantes lo recojan, si pueden llevarse (o no) el examen o una parte de él, cómo han de manejar los dispositivos electrónicos, si hay algún atajo de teclado que deban conocer, la forma en que han de iniciar sesión, etc.

- Se les indica que tienen un reloj que será visible durante toda la prueba.
- Si ningún candidato tiene ninguna duda en este punto, se procede a hacer una tercera comprobación del *hardware* y del *software* (en el caso de exámenes realizados por ordenador) y de los dispositivos de reproducción de audio. Esta tercera comprobación es particularmente importante para tranquilizar a los candidatos (que puede que estén viendo el ordenador de la prueba por primera vez o que tengan dudas sobre cómo escucharán el audio). Si todos los candidatos están satisfechos con lo anterior, se puede comenzar la prueba. En caso contrario, se harán los ajustes necesarios.

13. Al comenzar cada uno de los componentes de la prueba se ha de realizar un recuento de los candidatos que entran en la sala y, al terminar, se han de contar los exámenes recogidos. Ambas cifras han de coincidir entre sí y con el recuento anotado en el punto 11. Esto es particularmente importante si los candidatos tienen la posibilidad de abandonar la prueba al concluir alguna de las partes que la integran y antes de llegar al final.
14. Se ha de anotar e informar de cualquier incidencia acaecida (candidatos no presentados, nombres o identificaciones anotadas erróneamente en los listados, documentación sobrante o extraviada). Las incidencias más habituales deberían estar previstas en un protocolo de actuación que permita darles respuesta lo antes posible (por ejemplo, si se detecta que el nombre o identificación de un candidato está mal registrado, se ha de proceder a su modificación antes de la expedición de los certificados).

Después de la prueba

15. Se ha de comprobar que todos los exámenes quedan debidamente consignados (en soporte físico o digital), incluyendo las grabaciones de la producción oral en caso de que esta sea grabada.
16. Si se realiza corrección de «doble ciego», es decir, si dos correctores distintos han de corregir independientemente la misma prueba (por ejemplo, en el caso de producción escrita), estos han de recibir las diferentes copias de la prueba con suficiente antelación.
17. Las pruebas que no requieran corrección de «doble ciego» serán debidamente corregidas manual o automáticamente. En el caso de que la corrección sea automatizada, en ocasiones es posible que sea necesario que un corrector valide determinados ítems antes de que las hojas de respuesta se procesen.
18. Se aúnan todas las calificaciones (las procedentes de las pruebas de doble ciego, las procedentes de la corrección automática, etc.) y se aplican los

criterios de corrección establecidos (notas de corte establecidas durante el *standard setting*, calificación media global necesaria, etc.)

19. Se procede a comunicar a los candidatos sus calificaciones. Es recomendable realizar una primera publicación de calificaciones provisionales y dar margen para solicitar la revisión del examen, tras la cual se podrá realizar la publicación definitiva de las mismas.
20. Se procede a la emisión de los certificados correspondientes y al envío de estos a los candidatos.
21. Se ha de realizar un análisis estadístico de los resultados obtenidos y compararlos con lo descrito en la sección 3.5.5 para detectar posibles incoherencias.

Sobre esta base, las distintas instituciones habrán de establecer sus propios protocolos, que podrán modificar algunas de las secciones y añadir otras. Por ejemplo, puede ser interesante redactar instrucciones específicas para las pruebas orales, que pueden requerir una intendencia más compleja, ya que cuando se realizan de forma presencial suelen necesitar de varios tribunales.

3.5.9. Consideraciones posteriores a la prueba

Las pruebas no terminan en el momento en que los candidatos abandonan las instalaciones. A partir de ese momento, como hemos comentado en la sección anterior, tendrá que cerrarse todo el procedimiento de cara a los candidatos: será el momento de resolver las posibles incidencias detectadas, de corregir las pruebas según los criterios establecidos, de recibir las posibles reclamaciones de los candidatos, de emitir los certificados, etc.

Por otro, se deben realizar distintas consideraciones internas sobre la prueba. Por ejemplo, toda la información generada ha de ser debidamente consignada y almacenada el tiempo que establezca la normativa aplicable. Como también hemos comentado, es interesante repetir todos los análisis estadísticos descritos en la sección 3.5.5 pero, esta vez, con los datos de los candidatos reales. Esto nos permitirá tener más evidencias con las que construir nuestro argumento de validez, e incluso nos permitirá resolver posibles incoherencias de la prueba en caso de que decidamos reutilizarla en el futuro. En ocasiones, la normativa interna de las pruebas permite que, para ahorrar costes, un examen pueda reutilizarse total o parcialmente transcurrido un tiempo desde su último uso (por ejemplo, cada dos años) y, en cualquier caso, siempre que se tengan garantías de que los candidatos que concurrieron la primera vez al examen no volverán a concurrir a él fuera del ciclo de reutilización establecido.

CAPÍTULO 4
Diseño de Escalas

Se ha preservado un libro llamado el «libro estándar de escalas» […] que asigna valores numéricos al desempeño en las distintas asignaturas que son objeto de examen. Si así se requiriese, por ejemplo, se podría determinar el equivalente numérico de una muestra comparándola con cualquiera de las muestras estándar contenidas en el libro, que están numeradas en función de su valía, siendo las de mayor nivel las identificadas con el número 1, y las de menor nivel con el número 5. (Fisher, 1862:4)

Las páginas que siguen están dedicadas en exclusiva al diseño de escalas de producción, también conocidas como escalas evaluativas, escalas de corrección, escalas de dominio o rúbricas. La primera noticia que se tiene de un sistema de evaluación basado en escalas de este tipo data del siglo XIX. El explorador ártico y astrónomo George Fisher, con cuyas palabras abrimos este capítulo, asumió en la última etapa de su vida la dirección del Royal Hospital School de Greenwich, una prestigiosa academia naval británica en la que aún hoy se forman cientos de jóvenes año tras año. La formación científica de Fisher pronto le llevó a darse cuenta de que en el Royal Hospital School se carecía de un sistema de evaluación que reflejase de manera ajustada el desempeño de los cadetes, lo que generaba problemas a la hora de ubicar a los estudiantes en clases según su nivel. Para suplir esta carencia Fisher ideó lo que él mismo denominó «Standard Scale-book» (o libro estándar de escalas), en el que durante veinte años registró de forma sistemática las calificaciones de miles de cadetes mediante un sistema de puntuación de cinco niveles de desempeño (Fisher, 1862). Desde entonces, el uso de escalas evaluativas con distintos fines se ha extendido hasta convertirse en una parte fundamental de la evaluación de dominio de lenguas y, más concretamente, hasta convertirse en una parte central de la evaluación de las destrezas productivas.

Si un examen está bien diseñado, cualquier sistema automatizado puede aplicar con precisión una clave de corrección a las respuestas de los candidatos. Esto es posible debido a que en la corrección de ítems de respuesta múltiple, de emparejamiento, etc. no se establecen juicios o, mejor dicho, los juicios sobre las respuestas que serán consideradas como correctas se consensúan y revisan antes de que la prueba llegue a los candidatos. Incluso en aquellos casos en los que se realizan preguntas abiertas, la clave de la prueba debe reflejar con exactitud qué se considerará correcto y qué no. La corrección de una tarea de producción escrita u oral no puede, sin embargo, plantearse desde esta misma perspectiva dicotómica (correcto o incorrecto), ya que al analizar las respuestas de nuestros

candidatos tendremos que evaluar necesariamente las diferentes dimensiones de su desempeño (rango léxico, gramatical, pronunciación, fluidez, etc.).

La producción e interacción orales y escritas de nuestros candidatos son manifestaciones poliédricas y heterogéneas del lenguaje. En el caso de la producción oral, por ejemplo, es frecuente encontrarse con candidatos que, pese a no hablar de una manera muy fluida, dan muestras de tener un control preciso de determinadas estructuras sintácticas o palabras. En las pruebas de producción escrita, por otro lado, podemos encontrarnos con un candidato que, a pesar de mostrar cierta riqueza léxica, es incapaz de abordar el verdadero objetivo comunicativo de la tarea que se le ha requerido. Para reflejar todas estas diferencias dentro de una misma destreza lo adecuado es intentar descomponerlas en dimensiones (rango léxico, gramatical, pronunciación, fluidez, etc.) y establecer niveles de desempeño dentro de cada una de ellas. Para medir estos niveles de desempeño se diseñan las escalas de corrección, que son específicas de un constructo concreto y, por lo tanto, no intercambiables entre exámenes. Aplicar a una prueba las escalas que fueron diseñadas para otra sería lo mismo que intentar encajar piezas de dos puzles distintos.

En sentido estricto, dado que las escalas son una parte más de las pruebas de dominio, este capítulo bien podría haber sido un punto más en las secciones 3.4.4 o 3.4.5, en las que hemos hablado de las tareas de producción escrita y oral, o incluso en la sección 3.3, puesto que las escalas son un reflejo del constructo de nuestra prueba cuyo análisis contribuye a reforzar el carácter sustantivo del argumento de validez. No obstante, hemos decidido dedicarle un capítulo completo tras observar que el diseño de escalas ha estado tradicionalmente apartado de los manuales de diseño de exámenes de dominio publicados en español. La literatura sobre este aspecto es, por lo general, escasa y, con frecuencia, esotérica.

4.1. La forma y la función de unas escalas

En esencia, una escala está compuesta por descripciones de lo que diversos expertos consideran que es una destreza (normalmente producción e interacción orales y escritas). Las destrezas se componen de una o varias dimensiones[4] (una única dimensión en escalas holísticas y varias en escalas analíticas), dentro de las cuales se establecen diferentes niveles de desempeño organizados en bandas. Cada intersección entre dimensiones y bandas genera descriptores, breves

4. En lo sucesivo usaremos mayúscula inicial para hacer referencia a las dimensiones y subdimensiones de una escala con el objeto de distinguirlas de referencias genéricas a los elementos lingüísticos que las componen.

definiciones que delimitan con detalle los niveles observables en el continuo del dominio de cada dimensión.

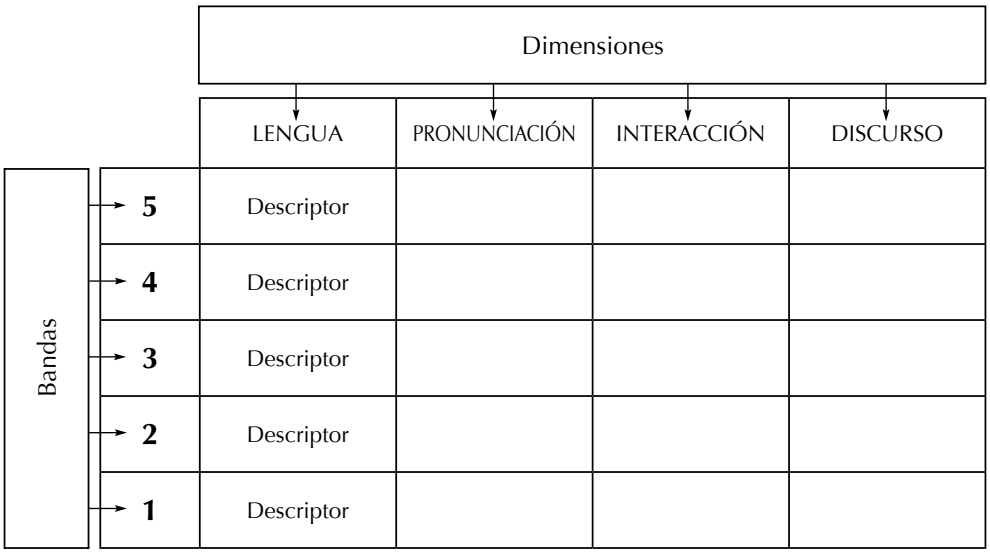


Figura 20. Disposición habitual de una escala analítica

En la figura 20 presentamos la forma típica de una escala analítica de producción oral dentro de la que, como vemos, se habrían establecido cuatro dimensiones y cinco bandas. Cada una de las dimensiones sería una de las facetas de las que se compone la competencia evaluada. Por ejemplo, en una escala analítica para producción oral se podrían encontrar Prosodia, Rango léxico, Sociopragmática e Interacción, si bien las dimensiones de las que se componga una escala dependerán del constructo de cada prueba. Cada una de las bandas es un peldaño dentro del continuo que supone el avance en el dominio de una lengua. Obviamente, los niveles que en nuestra escala aparecen en el eje horizontal, pueden ser ubicados en el vertical y viceversa. En una escala holística, como veremos más adelante, las distintas dimensiones se condensan en una descripción más general de la competencia de los candidatos.

Cualquier escala tiene el propósito de guiar a los evaluadores a la hora de asignar calificaciones a aspectos que requieren juicios subjetivos (Wolfe y Smith, 2007a:111). Tener unas escalas bien construidas permite reducir dicha subjetividad al mínimo. Como se observa en la figura 21 (Eberharter, comunicación personal), en estos juicios intervienen tres aspectos fundamentales: 1) la producción del candidato, 2) las escalas utilizadas y 3) el evaluador, que es quien filtra aquella en función de estas. De entre las tres variables (desempeño del

candidato, escalas de corrección y juicio del evaluador), la única que nunca se puede controlar es la producción del candidato. La actuación de los evaluadores, sin embargo, puede y ha de ser analizada regularmente para comprobar su consistencia (no es deseable que un candidato obtenga distintas calificaciones en función de si es evaluado por un calificador severo o benevolente). Las escalas también han de ser herramientas fiables que arrojen siempre los mismos resultados independientemente del evaluador que las maneje. Al pensar en una cinta métrica, por ejemplo, tenemos la certeza de que esta mide de la misma manera sin importar quién la utilice (un sastre, un albañil o un carpintero) e independientemente de qué objeto se mida (la manga de una chaqueta, la altura de una pared o el fondo de un cajón). Esta fiabilidad es la misma que deseamos que tengan nuestras escalas.

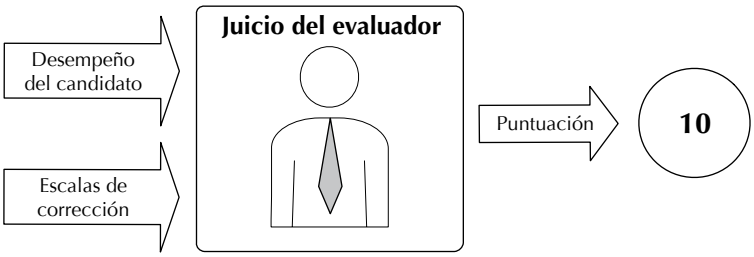


Figura 21. Relación entre candidato, escalas y evaluador

4.2. Cómo diseñar unas escalas

La literatura sobre el diseño de escalas es uno de los terrenos menos explorados en el ámbito de los exámenes de dominio en lengua española. Salvo algunas excepciones (North, 2000), la bibliografía sobre este tema (Wright y Masters, 1982; Consejo de Europa, 2001:207–212; Knoch, 2009; Knoch, 2011; Dean, 2012) es compleja o ambigua, está escrita en inglés, solo refleja una parte del proceso de diseño o es poco práctica. Debido a la dificultad que supone encontrar instrucciones claras sobre el diseño de escalas es frecuente que determinados exámenes de dominio hereden escalas de pruebas construidas con un constructo distinto, o que las diseñadas *ex novo* no se validen antes de ser usadas.

Proponemos la guía de más abajo al objeto de que quien desee crear una escala pueda proceder con método. La efectividad de esta guía ha sido probada en diferentes pruebas de dominio de impacto (Cruz, 2024), y en ella nos centraremos a lo largo de todo el capítulo 4:

Fase	Acción
1. Consideraciones iniciales	A. Elegir entre escala holística, analítica o de rasgos principales. B. Identificar el número de dimensiones y definirlas. C. Establecer el número de bandas. D. Establecer el formato de las calificaciones.
2. Confección de descriptores	A. Seleccionar las tablas del MCERVC con información relevante. B. Distribuir los descriptores del MCERVC como anclaje. C. Confeccionar descriptores de bandas incompletas. D. Revisar el conjunto de los descriptores.
3. Validación cualitativa	A. Revisión de expertos B. Modificar la escala en función de lo indicado por los expertos.
4. Validación cuantitativa	A. Al menos 2 evaluadores evalúan a 30 candidatos. B. Mediante Rasch se analiza la utilidad de las bandas, el *vertical ruler* y la fiabilidad de los evaluadores.
5. Implementación	A. Formar a los evaluadores en el uso de las nuevas escalas. B. Usar las escalas en condiciones reales de examen. C. Recoger y analizar los datos generados en el examen.
6. Revisión	A. Establecer un ciclo de revisión. B. Recoger los datos generados en diferentes convocatorias. C. Revisar los datos anteriores y repetir la fase 5 si es necesario.

Tabla 13. Guía para el diseño de escalas

4.2.1. Consideraciones iniciales

La primera fase de diseño es la más intuitiva y, al mismo tiempo, la que más debate suele generar (*vid*. Consejo de Europa, 2001:207–212). Este debate será beneficioso siempre que se aborde desde una perspectiva constructiva. Aunque puede que no siempre sea factible, en los compases iniciales del diseño de las escalas se debe intentar buscar consenso entre los confeccionadores de escalas y quienes luego hayan de utilizarlas, teniendo siempre como referente el constructo de la prueba. Al tener en cuenta a quienes usarán las escalas conseguiremos que la herramienta final no se perciba como la imposición de unos pocos sobre el trabajo de muchos, sino más bien como el fruto de un trabajo en el que todos estuvieron invitados a colaborar.

En ocasiones, sin embargo, lo mejor es enemigo de lo bueno, y si el debate inicial no se conduce de forma correcta, este puede producir un efecto contrario al deseado: descontrol, multitud de voces que defienden una postura y la contraria, y debates interminables que no alcanzan una conclusión. Es importante contar con un liderazgo sólido que supervise todas las fases descritas en la tabla 13. Este liderazgo ha de gozar tanto del reconocimiento académico de los compañeros como del respaldo administrativo necesario. Una buena idea para estas primeras fases de trabajo es pulsar la opinión de los distintos agentes interesados mediante una encuesta. Esto ayuda a hacer partícipe a un número extenso de colegas y, además, puede darnos pistas sobre el sentir general de usuarios finales. Cuanta más información aporte esta encuesta a los consultados (descripción del proyecto, punto de partida, objetivos perseguidos, plazos, etc.), mejor. Cuanta más información puedan aportar los encuestados (opinión sobre escalas preexistentes, preferencia sobre el número y la forma de las dimensiones, preferencia sobre el número de bandas, etc.), mejor.

Con los resultados de la encuesta en mente se puede ya crear un grupo de trabajo. En el caso de comunidades pequeñas (un departamento universitario de tamaño medio, por ejemplo), puede ser viable que todos sus integrantes participen en el debate sobre las consideraciones iniciales. En el caso de comunidades más extensas (comunidades educativas regionales o nacionales, por ejemplo), se debe convocar a un número de participantes operativo y representativo. Como decíamos anteriormente, que los colectivos implicados opinen en los primeros compases del diseño contribuirá a que las escalas finales no se perciban como una imposición sino como una decisión colegiada.

4.2.1.1. Elegir entre escala holística, analítica o de rasgos principales

La primera decisión importante que tendremos que tomar es la que define el tipo de escala que vamos a diseñar. Existen tres tipos principales de escalas usadas para la evaluación de lenguas, concretamente las:

- holísticas,
- analíticas y
- de rasgos principales.

Optar por una u otra dependerá a) de si la escala será usada para una tarea específica o para un conjunto de ellas y b) de si generará una o varias calificaciones (Weigle, 2002:109). Veamos qué caracteriza cada uno de estos tipos de escalas.

Las escalas holísticas (del griego *ὅλος*, pronunciado /ˈɔlɔs/, que significa «total», «completo») se usan para dar una puntuación global basada en una impresión general del desempeño de un candidato. Esta impresión general ha de estar basada en los descriptores que compongan la escalas, por lo que no se debe confundir una calificación basada en una impresión general con una calificación impresionista. Las escalas de este tipo pueden ir acompañadas de muestras de candidatos que ejemplifiquen los diferentes niveles descritos en ellas y que sirvan como referencia al evaluador, como ocurría en el «Standard Scale-book» de Fisher (1862) que mencionábamos al inicio del capítulo. Quienes defienden este tipo de escalas argumentan que son sencillas de manejar y que agilizan mucho el proceso de corrección, ya que contienen menos descriptores que las analíticas y, por lo tanto, suponen una carga cognitiva menor para los evaluadores. Los descriptores de las escalas holísticas suelen ser necesariamente más generales que los de las escalas analíticas. Si bien las escalas holísticas permiten a los evaluadores reaccionar de una manera más natural e intuitiva a la producción de los candidatos, pueden dar lugar a más subjetividad que la que generaría una escala analítica. El objetivo de nuestro diseño y de nuestra práctica debe ser reducir dicha subjetividad al mínimo y mantener el equilibrio entre funcionalidad y equidad.

La tabla 14 corresponde a la escala holística que el Instituto Cervantes utiliza para sus exámenes de dominio de español B2 DELE (Instituto Cervantes, 2014b:18). Como vemos, en esta escala, el dominio de la producción escrita del candidato no se divide explícitamente en dimensiones. Es recomendable, como hace esta escala, lograr que las diferentes bandas se distingan entre sí por definir distintos niveles de dominio de un mismo aspecto. En la tabla 14 observaremos que, por ejemplo, la «claridad del texto» y el «repertorio lingüístico» están presentes en todas las bandas. Las bandas superiores describen a candidatos con un dominio mayor que los que describen las bandas inferiores. En esta escala existe una banda 0. Probablemente esta banda (descrita en negativo, y no en positivo como sí lo están las anteriores) se incluyó para describir a aquellos candidatos que no consiguen alcanzar un mínimo. Conceptualmente, al menos dentro del sistema de educación y de enseñanza español, el 0 es el número que típicamente describe a los candidatos que no alcanzan un nivel mínimo. Técnicamente hablando, esta banda 0 debe ser considerada como una categoría más al analizar el funcionamiento interno de la escala, como veremos más adelante en la sección

4.2.4.2. El uso de las bandas 0 es controvertido ya que, si bien ayuda a identificar a determinados grupos de candidatos, estos grupos suelen ser muy reducidos y, por lo tanto, dicha banda se utiliza con poca frecuencia, lo que puede llevar a problemas de representatividad durante el análisis estadístico de las escalas.

DELE B2. Expresión e interacción escritas - Escala holística	
3	Proporciona los contenidos más importantes del texto de entrada y construye con eficacia los argumentos. Produce textos claros, bien estructurados y desarrolla todos los puntos de orientación. Elige los elementos lingüísticos más apropiados para la construcción del discurso y muestra un alto grado de corrección gramatical que le permite expresarse con claridad, aunque pueda producirse esporádicamente algún error, desliz o imprecisión gramatical, estructural o léxica.
2	Produce textos suficientemente claros, elaborando argumentos y la información de otras fuentes, aunque puede producirse alguna vacilación en el desarrollo o en la estructura. Utiliza un repertorio lingüístico amplio y suficiente para completar las tareas aunque puede producirse algún error de escasa entidad.
1	El texto responde a la situación, pero la sencillez, brevedad o falta de claridad hacen que resulte inapropiado para conseguir el objetivo. Muestra un repertorio lingüístico limitado compuesto por estructuras muy sencillas o con errores léxicos o sintácticos elementales que dificultan la comprensión y que, en algunos casos, no dejan clara la idea general.
0	El texto no responde a la situación y no consigue el objetivo comunicativo planteado por su inadecuado desarrollo, su desorganización o por la presencia de abundantes errores que dificultan la comprensión del mensaje.

Tabla 14. Escala holística para el nivel B2 del examen DELE

Existen otros tipos de escalas holísticas que, además, ofrecen la posibilidad de puntuar subjetivamente dentro de un rango establecido para una banda concreta (*vid*. Dean, 2012:42–46). En estas escalas, el evaluador primero determina la banda global en la que considera que el candidato ha de estar ubicado. Después, dentro de cada una de las bandas existe una horquilla de puntos que el evaluador puede otorgar a cada candidato a discreción, sin que se establezca qué criterios determinan una u otra puntuación dentro de la mencionada horquilla. En nuestra opinión, el uso de puntuaciones no adscritas a un descriptor concreto añade subjetividad a la escala. Si no existe un descriptor que diferencie lo que es un 1 de lo que es un 5, por ejemplo, se corre el riesgo de que diferentes evaluadores ofrezcan calificaciones impresionistas. Pudiera parecer que las puntuaciones no adscritas a

descriptores facilitan el trabajo, pero, en realidad, si no se realiza un seguimiento estricto de la fiabilidad de los evaluadores, tan solo se está generando una ficticia sensación de agilidad que casi siempre desemboca en aleatoriedad. Analizaremos un ejemplo de esto en la sección 4.2.1.4.

Las escalas analíticas, por su parte, son aquellas que distinguen varias dimensiones dentro de una destreza. Estas escalas generarán una puntuación por cada una de dichas dimensiones que, a su vez, podrán convertirse en una puntuación global. Quienes defienden este tipo de escalas lo hacen señalando que, dado que contienen definiciones más específicas que las escalas holísticas, reflejan mejor el desempeño de un candidato, que no siempre tiene que ser el mismo en todas las dimensiones, de igual manera que no suele ser uniforme entre destrezas.

En la tabla 15 observamos un ejemplo de escala analítica, concretamente la escala para la prueba de expresión e interacción escritas del examen del SIELE (Servicio internacional de evaluación de la lengua española, 2021:47). En esta ocasión, al tratarse de una prueba multinivel, la escala no hace referencia a un nivel concreto, al contrario de lo que ocurría en la escala de la tabla 14, que era exclusiva del nivel B2. Como vemos, la escala del SIELE distingue hasta tres dimensiones, concretamente Cohesión, Corrección y Alcance. La escala de este examen incluye una cuarta dimensión, Cumplimiento de la tarea, que es distinta para cada una de las dos tareas que componen la prueba y que, por cuestiones de espacio, no reproducimos aquí.

ESCALA DE CALIFICACIÓN			
PARA LAS DOS TAREAS			
Puntos	COHESIÓN	CORRECCIÓN	ALCANCE
5	Produce un texto claro, coherente y muy bien estructurado, en el que demuestra un uso bastante completo y variado de estructuras organizativas, de una amplia serie de conectores para marcar la continuidad o los cambios de tema, y de otros mecanismos de cohesión (subordinadas sustantivas, adjetivas y adverbiales; recursos de referencia como deícticos…). Utiliza correctamente las reglas de puntuación.	Muestra un consistente dominio de un nivel de lengua complejo, sin apenas errores de léxico, gramática u ortografía.	Describe, narra o argumenta en todo tipo de situaciones, sin tener que restringir lo que quiere decir; es capaz de detallar y concretar los temas de los que trata, aunque sean abstractos; y sabe defender sus puntos de vista con argumentos y ejemplos, así como elaborar conclusiones. Sus textos son claros, ricos y precisos, tienen un estilo personal, apropiado para el lector al que van dirigidos, y pueden incluir expresiones idiomáticas y coloquiales en el contexto adecuado.

4	Produce un texto claro, coherente y estructurado. Para enlazar las frases hace un uso relativamente variado de mecanismos de cohesión, como conectores (a pesar de, por lo tanto, no solo… sino también), organizadores de la información (para empezar, finalmente, por otra parte, en cuanto a), recursos de referencia (marcadores, deícticos, etc.) y la subordinación de oraciones. No obstante, puede haber algún error esporádico en las referencias y los conectores o poca claridad en una frase o en la relación entre dos frases o partes del texto. Utiliza correctamente las reglas básicas de puntuación.	Mantiene un buen control gramatical (uso de los pasados, perífrasis verbales, subordinadas con subjuntivo, uso de lo, comparativos…) y léxico, aunque todavía puede cometer algunos fallos en la estructura de oraciones largas o complejas (sentimientos, temas abstractos…) o en el vocabulario específico o menos frecuente. La ortografía es razonablemente correcta, pero puede cometer algún error en los acentos y en el léxico menos habitual.	Es capaz de explicar los puntos principales de una idea o un problema, de expresar sentimientos y pensamientos sobre temas complejos o abstractos, de sintetizar y evaluar información y argumentos, y de describir situaciones no habituales. Se expresa de forma sencilla pero razonablemente precisa, para lo que puede incluir algunas expresiones idiomáticas y coloquiales.
3	Escribe textos cohesionados, ordenados mediante una secuencia lineal de elementos sencillos, utilizando organizadores de la información (primero, luego, después), conectores frecuentes (también, entonces, porque, así que, además, aunque, sin embargo…), relativos (donde, cuando…), subordinadas sustantivas (creo que, …), aunque el texto puede presentar alguna deficiencia o limitación en la relación entre sus partes o en el uso de mecanismos de cohesión. Utiliza correctamente las reglas básicas de puntuación.	Muestra un control razonable de elementos lingüísticos sencillos y estructuras habituales (distinción ser y estar en sus usos más comunes, imperativo, presentes irregulares, uso básico de los pasados, uso de las perífrasis verbales más frecuentes…) en temas predecibles y generales, pero comete errores gramaticales, imprecisiones léxicas y/o fallos ortográficos.	Es capaz de solicitar información, hacer valoraciones, expresar deseos, dar instrucciones, hacer descripciones sencillas pero adecuadas de temas cotidianos, hechos y viajes, o narrar historias. La falta de especificidad y los circunloquios muestran sus limitaciones, y puede cometer imprecisiones, repeticiones o errores si se expresa sobre temas complejos o abstractos.

2	Escribe textos básicos con oraciones breves enlazadas mediante recursos algo limitados: conectores sencillos (y, pero, porque, por eso…), relativos (que), pronombres… Pueden producirse errores (uso indebido de elementos de referencia, elección indebida de deícticos, falta de organización del texto) y fallos o imprecisiones en la puntuación que dificulten la lectura, aunque no afectan al significado.	Muestra un control elemental de elementos lingüísticos sencillos, pero sistemáticamente comete errores básicos de gramática (ser/estar/ haber, formas de los tiempos verbales regulares e irregulares, pronombres, concordancias de sujeto, verbo o nombre-adyacentes…), de léxico y de ortografía, aunque se entiende lo que quiere transmitir.	Es capaz de transmitir información básica en situaciones concretas y cotidianas (dar información personal, hablar de aspectos del entorno más cercano, etc.) y expresarse sobre temas conocidos, sencillos y habituales (describir de forma breve y básica acciones, experiencias personales o situaciones frecuentes, solicitar información puntual, etc.).
1	Escribe una serie de frases sencillas o grupos de palabras enlazados con conectores muy básicos (y, pero). El discurso no mantiene una estructura organizada y la información aparece desordenada. Hay errores de puntuación.	Utiliza estructuras gramaticales muy básicas y sencillas relacionadas con necesidades básicas inmediatas. Comete abundantes errores gramaticales (concordancias, personas del verbo, formas del presente…), léxicos y ortográficos, lo que dificulta la comprensión del mensaje.	Su nivel de dominio se limita a los datos personales y las necesidades inmediatas, aunque puede que cumpla otras funciones comunicativas utilizando recursos de lenguas cercanas.
0	No elabora frases completas. Escribe palabras sueltas sin coherencia entre ellas.	Apenas utiliza estructuras gramaticales y sintácticas o léxico correctos.	Utiliza algunas palabras en español, pero apenas es capaz de expresarse.

Tabla 15. Escala analítica multinivel del examen SIELE

Por último, las escalas de rasgos principales están orientadas a definir el nivel de desempeño en un aspecto específico del discurso que es especialmente relevante en nuestro constructo, por ejemplo, cuán hábil es un candidato escribiendo textos persuasivos o explicativos, o su capacidad de transmitir oralmente mensajes claros y concisos. Debido a su especificidad (una escala de rasgos principales se diseña exclusivamente para una tarea en concreto), estas escalas tienen una aplicación muy limitada y no son prácticas (Knoch, 2011:83). Según Weigle (2002:110),

estas escalas están compuestas por a) la propia tarea; b) una indicación sobre el aspecto específico del discurso que se quiere analizar en el candidato (por ejemplo, la capacidad de persuasión en una reseña, la habilidad descriptiva en una narración, la capacidad de transmitir mensajes claros y concisos, etc.); c) una hipótesis sobre el desempeño esperado en la tarea; d) una explicación de cómo el rasgo principal se vincula con la tarea; e) una escala que articule los niveles de desempeño; f) ejemplos de desempeño de los diferentes niveles y, en cada una de las correcciones que se lleven a cabo mediante lo anterior g) una explicación de por qué cada candidato ha sido calificado de una forma concreta. Se pueden encontrar ejemplos de escalas de rasgo principal en Weigle (*ibid.*:111).

En la figura 22 adaptamos un ejemplo de escala de rasgos principales tomado de Lloyd-Jones (1977). El ejemplo reproducido incluye la tarea, indicaciones sobre el aspecto discursivo que se desea analizar y la escala propiamente dicha. No incluye, sin embargo, la hipótesis sobre el desempeño esperado, ejemplos de respuestas tipo ni correcciones argumentadas. Estas hipótesis, ejemplos y calificaciones, que se extienden a lo largo de 13 páginas, pueden consultarse en el original (*ibid.*).

Como queda evidenciado, la construcción de una escala de este tipo es costosísima en tiempo si tenemos en cuenta que son específicas para una tarea exclusivamente. Según Lloyd-Jones (1977:44–45), puede conllevar hasta ochenta horas de trabajo. Además, su uso supone una tremenda carga cognitiva para el calificador (pensemos en corregir mientras se tienen presentes las mencionadas trece páginas de hipótesis y ejemplos), por lo que su uso es difícilmente justificable.

Observe con atención esta fotografía. Estas personas están divirtiéndose en la playa. **Escríbale** a un amigo explicando lo que ocurre en la fotografía como si fuera usted un **protagonista** o un **observador**. Transmita **sentimientos vívidos** para contribuir a que su amigo **sienta la experiencia** igual que usted.

Contexto
Rasgo principal que se ha de analizar: expresión de sentimientos vívidos a través de la elaboración imaginativa de un punto de vista concreto.

Escala de puntuación

Calificación global
0. No hay respuesta. Oraciones fragmentadas.
1. Responde parcialmente.
2. Responde a lo requerido.

Uso del diálogo
0. No usa diálogos en su respuesta.
1. Utiliza el estilo directo con una persona en una o varias intervenciones.
2. Usa el estilo directo con dos o más personas de la historia.

Punto de vista
0. No se puede determinar un punto de vista en la historia.
1. Expresa el punto de vista de uno de los protagonistas.
2. Expresa el punto de vista como si fuera un observador.

Tiempo verbal
0. No usa los tiempos verbales de forma consistente.
1. Usa el presente.
2. Usa el pasado.
3. Usa subjuntivo.

Figura 22. Escala de rasgos principales

Es frecuente encontrar exámenes de dominio que combinan el uso de varios tipos de escalas. Por ejemplo, en el caso de los exámenes SIELE y en algunos niveles del DELE se usa una combinación de escalas analíticas y holísticas. En el caso de los exámenes DELE de nivel A2, además, a partir de 2020, la combinación de escalas analíticas y holísticas se revisó y ambas se refundieron en una escala híbrida para este nivel. Una combinación de escalas holísticas y analíticas puede ser una opción interesante. Imaginemos que, en una prueba de producción oral, además del candidato, participan dos evaluadores, uno de los cuales ha de conducir la interacción mientras que el otro solamente escucha. En estos casos, el evaluador que dirige la conversación tendrá más libertad si utiliza una escala holística mientras que el otro corrector usa la escala analítica para captar matices más sutiles. Esto no sería necesario en las escalas para la corrección de producción escrita, dado que estas se corrigen tras la prueba y sin que el evaluador tenga que gestionar ninguna interacción con los candidatos. En los casos en que se combinen dos tipos de escalas será importante determinar de forma clara el peso que cada una tiene en la puntuación final. En el caso del DELE descrito anteriormente, por ejemplo, la puntuación de la escala holística supone un 40 % de la calificación final del candidato mientras que la analítica contribuye con el 60 % restante (Instituto Cervantes, 2014b:13, 27). En el caso del examen SIELE, su guía menciona «una fórmula específica para cada prueba» de la que no dan más detalles (Servicio Internacional de Evaluación de la Lengua Española, 2021:46).

Elegir uno u otro tipo de escalas depende en gran medida de la tradición, del gusto y de las necesidades de quienes tengan que utilizarlas. No hay un modelo superior a otro, tan solo uno que se adapta mejor a nuestras necesidades. Algunas de las consideraciones que se suelen tener en cuenta a la hora de elegir el modelo de escala son su especificidad o su versatilidad para diferentes tipos de tareas. Por ejemplo, se suele pensar que las escalas analíticas analizan mejor el dominio de un candidato pero que, al mismo tiempo, nos pueden privar de una visión de conjunto del desempeño. Unas escalas holísticas, por el contrario, ofrecen esa visión de conjunto a costa de especificidad. Tengamos en cuenta, no obstante, que es poco práctico desarrollar una escala tan específica que contenga demasiados descriptores o descriptores muy largos. Excesos de este tipo sobrecargan la memoria de trabajo del evaluador durante la corrección. Como vimos en la figura 21, en la calificación final de nuestros candidatos convergen su desempeño, nuestras escalas y también la capacidad cognitiva de los evaluadores. Una escala mal diseñada aumentará la fatiga del evaluador y por ello hemos de diseñar herramientas livianas, sencillas de entender y fáciles de usar.

Es evidente que, cuanto más centrada esté una escala en un tipo de tarea, más específica podremos hacerla. Por el contrario, si buscamos crear una herramienta aplicable a varias tareas, tal vez estaremos comprometiendo esta especificidad para ganar en agilidad. Imaginemos que nuestro examen de dominio tiene una sección de producción escrita en la que los candidatos han de realizar dos

tareas (dos composiciones escritas de un nivel determinado) distintas entre sí y cada una con características muy concretas. Supongamos que la tarea 1 consiste en redactar un ensayo de opinión, mientras que la segunda requiere escribir un correo electrónico. Es fácil intuir que las destrezas necesarias para escribir una y otra son diferentes. Podríamos, dadas estas especificaciones, construir dos escalas de rasgos principales, una para cada una de las tareas, o bien construir una escala menos encasillada (holística o analítica) que nos sirviera para ambas tareas. A mayor número de escalas, mayor será el esfuerzo para los correctores. Desde otra perspectiva, cuanto más específicas sean las escalas, más exacta podrá ser, por ejemplo, la información que se les dé a los candidatos tras la prueba. Como vemos, toda decisión supone una concesión. Con nuestro constructo como referencia, hemos de crear la herramienta de medida que mejor se adapte a nuestras pruebas teniendo en cuenta tanto a los candidatos como a los correctores.

4.2.1.2. Identificar el número de dimensiones y definirlas

Una vez que hemos decidido qué tipo de escala usaremos (analítica, holística o de rasgos principales) tendremos que determinar qué aspectos concretos del desempeño de nuestros candidatos queremos analizar. Entendemos por dimensiones todos aquellos componentes que, en su conjunto, forman la destreza que deseamos evaluar según esta está definida en nuestro constructo. Imaginemos, por ejemplo, que queremos diseñar una escala para producción oral. Ejemplos de dimensiones dentro de la producción oral serían Pronunciación, Fluidez, Prosodia, Lengua, Corrección, Alcance, Discurso, Interacción, etc. Si estuviésemos diseñando una escala que midiese la producción escrita hablaríamos de dimensiones como Lengua, Coherencia, Cohesión, Ortografía, etc.

La elección de dimensiones suele generar debate. Si el constructo de nuestra prueba no está firmemente interiorizado por los miembros del equipo de confección, lo más probable es que sus componentes intenten reflejar en dichas escalas una visión particular de la lengua. Así, por ejemplo, habrá quienes defiendan que exista una dimensión denominada Lengua, que se defina como la suma de gramática y vocabulario, mientras que otros defenderán que Gramática y Vocabulario deben ser consideradas como dimensiones independientes; habrá quienes aboguen por el uso de una dimensión centrada en la Sociopragmática; en la producción oral habrá quienes defiendan que la Pronunciación tenga el mismo peso que la Interacción y habrá quienes defiendan lo contrario, etc. Por mucho que todos los componentes del equipo defiendan diferentes posturas con vehemencia y con argumentos válidos, hemos de recordar que entrar en estas discusiones puede ser enriquecedor hasta cierto punto pero que, si las discusiones se mantienen en el tiempo sin un objetivo concreto, harán que las posturas se enroquen, generando así un efecto opuesto al consenso deseado. Es útil recordar de nuevo que, como decíamos unas líneas más arriba, las escalas han de reflejar el constructo de nues-

tra prueba y adaptarse a nuestras especificaciones. Las dimensiones de nuestras escalas no tienen que ser un trasunto de nuestra concepción individual de la lengua, ya que, si así fuese, existirían tantas escalas como confeccionadores hubiera en el equipo de diseño, algo que, evidentemente, no es viable.

Para que nuestras escalas sean sencillas de usar es interesante definir en algún lugar visible de las mismas las dimensiones que finalmente se escojan. Si nuestra escala analítica de producción oral está compuesta, por ejemplo, por Lengua, Discurso, Pronunciación e Interacción, debemos definir qué significan estas dimensiones antes incluso de confeccionar los descriptores. Tener claras estas definiciones nos ayudará a mantener un hilo conductor cuando redactemos dichos descriptores. A modo de ejemplo, la definición de estas dimensiones podría quedar de la siguiente manera:

- Lengua: compuesta por el alcance y control del candidato en vocabulario y gramática. Además, tendrá en cuenta los errores que se cometan.
- Discurso: analizará el desarrollo temático que realice el candidato y la cohesión y fluidez con que lo haga.
- Pronunciación: analizará la pronunciación general, la articulación de los sonidos y la prosodia (acento, ritmo y entonación) del candidato.
- Interacción: tendrá en cuenta la efectividad con que se realiza el intercambio de información necesaria, la destreza con que el candidato inicie y mantenga la conversación y su capacidad para cooperar y llegar a un punto de entendimiento.

Otras dimensiones y definiciones son posibles. Lo que es interesante de estos ejemplos es el hecho de que cada una de las definiciones identifica tres subdimensiones:

- Lengua = Vocabulario + Gramática + Errores
- Discurso = Desarrollo temático + Cohesión + Fluidez
- Pronunciación = Pronunciación general + Articulación de sonidos + Prosodia
- Interacción = Intercambio de información + Conversación + Cooperación

Tener claramente identificadas las subdimensiones será útil, ya que, más adelante, cada una de las bandas que confeccionemos deberá contener al menos una referencia a cada subdimensión, lo que nos permitirá reflejar diferentes niveles de destreza de una forma progresiva a lo largo de todas las bandas. No es recomendable, por ejemplo, definir la banda 1 de una escala de acuerdo a unos subcomponentes y definir la banda 5 de acuerdo con otros distintos, ya que, al usar las escalas, el corrector observará que, en algún momento, la definición de la dimensión cambia y esto puede generar confusión. De igual manera es importante buscar descripciones que puedan reflejarse en todas las bandas. Dotar a nuestras

escalas de esta continuidad a lo largo de las diferentes bandas no siempre será sencillo. Por ejemplo, se pueden dar problemas al diseñar una escala analítica multinivel (A2–C2) para producción escrita que trate de incluir el Registro como una de sus subdimensiones ya que, si bien este se puede observar con claridad en niveles altos (C1, C2), es más difícil apreciarlo en niveles bajos. Lo mismo puede ocurrir con otras subdimensiones.

Sabemos que las competencias que nos ayudan a articular el lenguaje no se manifiestan aisladas unas de otras. De igual manera que, por ejemplo, la competencia gramatical está presente tanto en la producción oral como escrita, las escalas que usemos para medir una y otra podrán compartir características. Cualquier dimensión que identifiquemos y definamos es susceptible de ser utilizada en otra destreza dentro de la misma prueba. Esto agiliza el uso de las escalas, ya que los correctores reconocerán dichas dimensiones.

En las escalas holísticas no existen dimensiones propiamente dichas, dado que lo que se busca es reflejar descripciones globales del desempeño de los candidatos. Lo que encontramos en las escalas holísticas es una única descripción del desempeño de los candidatos que engloba varios aspectos, como se observa en la tabla 14 de la sección 4.2.1.1.

Si hemos optado por una escala de rasgos principales, probablemente los aspectos que hemos de evaluar estarán ya claros en este punto porque, recordemos, fue la relevancia de dichos rasgos lo que nos condujo a decantarnos por ellas. Estos rasgos principales coparán toda nuestra atención y confeccionaremos nuestros descriptores en torno a ellos (*vid.* Weigle, 2002:111).

4.2.1.3. Establecer el número de bandas

Si las dimensiones reflejan los componentes que distinguimos dentro de una destreza, las bandas establecen los diferentes niveles que el candidato puede alcanzar en el continuo de cada dimensión. Es útil imaginar las bandas como los escalones que nos elevan desde la parte baja de una dimensión a la parte alta, y que equivalen a diferentes etapas del dominio de una lengua. Es necesario concebir cada uno de estos escalones como una categoría, una especie de compartimento estanco, bien delimitado, en el que colocaremos al candidato en función de lo que observemos en su desempeño. Cuantas más bandas establezcamos, mejor podremos definir diferentes estadios de progresión. No obstante, esta mayor concreción también dota a las escalas de más contenido, un contenido que los evaluadores tendrán que utilizar constantemente y que, por lo tanto, incrementará el esfuerzo cognitivo necesario.

La sobrecarga cognitiva de unas escalas y el esfuerzo que supone un número excesivo de bandas para la memoria de trabajo de un evaluador (Richards y Schmidt, 2002; Baddeley, 1988; 2000a; 2000b; 2003) son algunas de las razones que en ocasiones llevan a los confeccionadores de escalas a no definir mediante descriptores determinadas bandas intermedias, y a considerarlas como bandas que

comparten características de las adyacentes. Esta suele ser una decisión controvertida, ya que, si bien eliminar los descriptores de una banda flexibiliza el uso de la escala en su conjunto, también puede hacer que los correctores identifiquen esas bandas como bandas vacías de contenido. Se corre el riesgo de usar las bandas en blanco como un cajón de sastre en el que ubicar todo aquello que no encaja de forma clara en otras bandas. Usar las bandas de forma aleatoria compromete la fiabilidad de la escala y, por ello, como veremos más adelante en la sección 4.2.4.2, durante el análisis cuantitativo de nuestras escalas será necesario analizar con detalle si la cantidad de veces que los evaluadores usan cada banda es regular o si las mediciones medias avanzan monótonamente en cada una de ellas. Es decir, habrá que observar si se eligen con la misma frecuencia las bandas que están en blanco y las que no, y si el esfuerzo que los candidatos han de realizar a la hora de subir los diferentes escalones es siempre el mismo.

El número de bandas de una escala y su definición también están ligados al constructo de una prueba y a los niveles que esta evalúa. En exámenes multinivel la escala tendrá que abarcar un rango de niveles más amplio que el definido por escalas de exámenes de un solo nivel. Las escalas del examen multinivel IELTS, por ejemplo, contienen 10 bandas (numeradas de 0 a 9) que describen todos los posibles niveles de dominio existentes entre A1 y C2. Las escalas del examen *uninivel* CPE contienen seis bandas (numeradas de 0 a 5) que definen el nivel C2. Si aplicásemos a una escala multinivel A1–C2 el mismo grado de concreción que usan las escalas *uninivel* CPE, obtendríamos una escala con treinta bandas, algo a todas luces impracticable.

Se ha de armonizar, pues, el nivel de concreción y la carga cognitiva que suponen las escalas. La carga cognitiva de una escala analítica de cinco bandas y cuatro dimensiones es mayor que la de una escala holística de diez bandas. Un número impar de bandas mayor de tres nos da la oportunidad de ubicar en la banda intermedia el aprobado. Para evitar la tendencia central es recomendable no usar escalas con solo tres bandas. Una escala de tres bandas puede convertir las calificaciones en dicotómicas, es decir, calificaciones del tipo «sí tiene el nivel» (y por lo tanto ubico al candidato en la banda 2 o 3) frente a «no tiene el nivel» (y, por lo tanto, lo ubico en la banda 1). El siguiente número impar de bandas sería el cinco, un número versátil en escalas de todo tipo. Cinco es también el número de bandas sugerido por el Consejo de Europa (2001:181–182), que asimismo recomienda ubicar en la banda 3 el aprobado y extraer el resto de descriptores de los niveles adyacentes.

Este modelo de escalas de cinco bandas (escala tipo A de la figura 23) es práctico por su sencillez conceptual, por el tipo de errores que puede ayudarnos a evitar y por su escalabilidad. Por un lado, nos ayuda a entender que el aprobado de nuestra escala es la banda intermedia, que los candidatos con mejor dominio estarán por encima, y que los candidatos menos hábiles quedarán por debajo. También nos ayuda a evitar posibles errores de diseño que pueden generarse si se excluyen de la escala los niveles adyacentes al aprobado (escala B1 de tipo B en

la figura 23) o si se ubica el aprobado en lugares incorrectos de la escala (escala B1 de tipo C en la figura 23). El problema de un diseño de tipo B es que se trunca artificialmente el continuo del dominio de una lengua y se fuerza la redacción de descriptores adicionales para dotar de contenido a las bandas 1, 2, 4 y 5. El problema de una escala de tipo C es que se definen en exceso los niveles inferiores al aprobado (que se mantiene en la banda 5) y no se define ningún nivel por encima de este, por lo que la herramienta carecerá de sensibilidad para reflejar el dominio de candidatos con un nivel superior al aprobado.

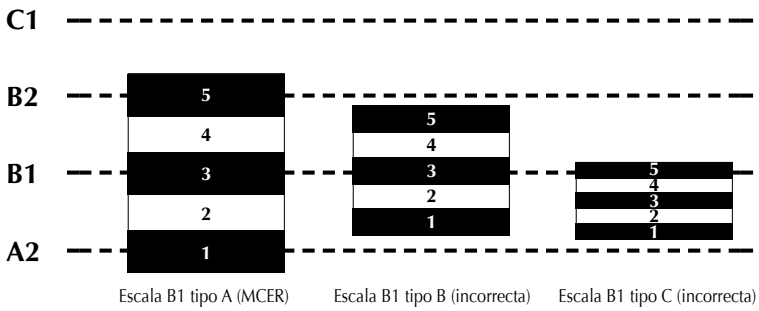


Figura 23. Escalas según la ubicación del aprobado y los niveles adyacentes

El del Consejo de Europa (2001:181–182) es, además, un modelo escalable que nos permite reciclar descriptores entre escalas de diferentes niveles. Como se observa en la figura 24, los niveles 3, 4 y 5 de la escala A1 son los mismos que los niveles 1, 2 y 3 de la escala A2, etc., algo que sería imposible si se hubiese optado por escalas de tipo B o de tipo C.

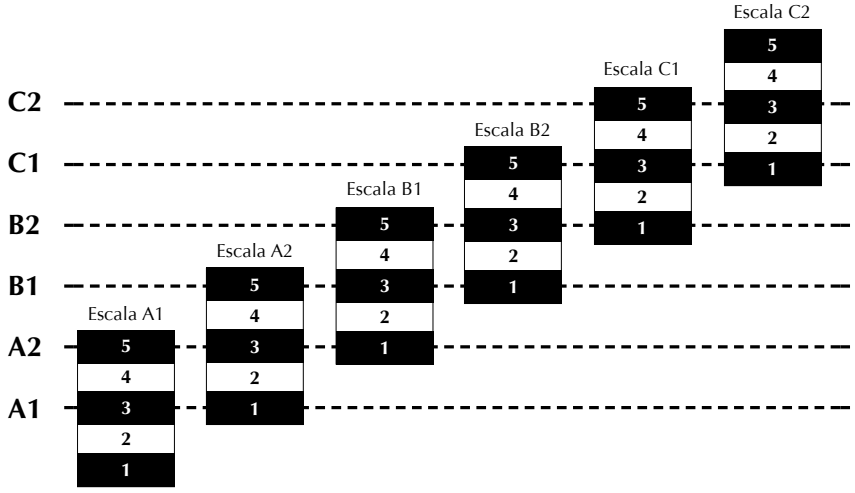


Figura 24. Modelo escalable de diseño

4.2.1.4. Establecer el formato de las calificaciones

Cada examen tiene su propio sistema de calificación. El examen de inglés TOEFL iBT, por ejemplo, establece una puntuación que oscila entre 0 y 120; en IELTS las calificaciones pueden ir desde 0 a 9 y ascienden de medio en medio punto; los exámenes SIELE de español establecen una puntación que va desde 0 a 250; el examen de francés TCF otorga puntuaciones entre 0 y 699, y en otros tantos casos la calificación es sencillamente apto o no apto.

Al diseñar cualquier escala de dominio es importante considerar el sistema de calificación de la prueba a la que sirven. Imaginemos que estamos diseñando unas escalas para un examen de cuatro destrezas en el que la puntuación final del candidato, sumados estos componentes, será un número entre 0 y 100. Para dicha escala elegimos el modelo de cinco bandas y cuatro dimensiones con el aprobado en la banda 3. En esta escala, el máximo de puntuación que un candidato puede obtener es 20 (5x4). Si estimamos que el peso de la producción escrita es 1/4 de la calificación global (producción oral, comprensión lectora y comprensión auditiva serían los 3/4 restantes), un 20 en la producción escrita equivale a 25 puntos en la calificación global. Se trata de una simple regla de tres aplicable a cualquier modelo de puntuación.

Además de considerar cómo se relacionan las puntuaciones de las escalas con la prueba en su conjunto, hemos de analizar también las mecánicas internas de puntuación de la propia escala. En el caso de la hipotética escala antes mencionada de cuatro dimensiones y cinco bandas, con el aprobado en la banda 3, el candidato que obtiene una puntuación de 3-3-3-3 consigue 12 puntos de un máximo de 20 posibles. Por un lado, vemos que 3-3-3-3 es una puntuación que claramente identifica al candidato como aprobado pero, por otro, vemos que 12 puntos no es la mitad de 20. Dicho de otra manera, lo que es un aprobado en las escalas es más del 50 % de la puntuación si se utiliza una regla de tres. Dependiendo del constructo de la prueba, una puntuación de 3-3-2-2 también podría considerarse como aprobado, dado que la competencia lingüística, aun cuando se considere dentro de una misma destreza, no es siempre pareja en todas sus dimensiones. Esto es lo que el Consejo de Europa (2020:38–40) denomina perfiles, es decir, diferentes niveles de evolución en las distintas destrezas o dimensiones de la lengua. Podemos encontrar candidatos que dominen un abanico extenso de vocabulario, pero cuyo repertorio sintáctico sea reducido, personas que hablen con mucha fluidez a pesar de tener una pronunciación con marcada influencia de la lengua materna, etc.

En una distribución tradicional de cuatro destrezas en las que todas ellas aportan lo mismo, la producción oral y escrita suponen la mitad de la prueba, pero no siempre tiene que ser este el caso. En un hipotético examen de nivel para controladores aéreos, por ejemplo, podría resultar necesario que el componen-

te oral de la prueba supusiese 4/5 de la calificación final. Estas consideraciones (peso de los diferentes componentes, fórmulas mediante las que se calcula la calificación final, etc.) han de estar reflejadas en las especificaciones y deben emanar del constructo de la prueba.

Como ya hemos mencionado anteriormente, las bandas son categorías, es decir, compartimentos estancos bien delimitados en los que ubicaremos a los candidatos en función de su desempeño. Por este motivo creemos que es importante huir de puntuaciones fraccionadas (1.25; 3.8; 7.50, etc.) en el diseño de las bandas. La experiencia nos dice que este tipo de puntuaciones pueden llevar a malinterpretaciones conceptuales y a la subjetividad. Si una escala contiene cinco bandas (es decir, cinco niveles de dominio bien diferenciados), lo óptimo sería numerar dichas bandas como 1, 2, 3, 4 y 5 y no como 1; 1.5; 2; 2.5 y 3 o como 0, 1, 2, 3 y 4.

En principio, desaconsejamos modelos como el descrito por Dean (2012:42-46), ya mencionado en la sección 4.2.1.1, en el que existe una horquilla de puntos que el evaluador puede otorgar a cada candidato a discreción, sin que se establezca qué criterios determinan una u otra calificación dentro de la mencionada horquilla. El motivo para desaconsejar este tipo de escalas es, de nuevo, la posible aleatoriedad que se deriva de que una puntuación no esté ligada a un criterio específico. Hemos encontrado casos, no obstante, en los que establecer esta horquilla no influía de manera negativa en las escalas. Por ejemplo, en las escalas de la Unidad de Evaluación y Certificación en Lenguas de la Escuela de Idiomas de la Universidad de Antioquia, Colombia. Estas escalas contienen diferentes horquillas de puntuación que no son similares entre bandas ni entre dimensiones. La escala de producción oral está dividida en cuatro dimensiones (Competencia Pragmática, Competencia Gramatical, Competencia Léxica y Competencia Fonológica) y seis bandas (A1-C2). El rango de puntuación de la Competencia Pragmática para las seis bandas oscila entre los 0 y los 35 puntos (0–10.4; 10.5–17.4; 17.5–24.4; 24.5–27.9; 28–31.4; 31.5–35), mientras que el de la Competencia Gramatical oscila entre los 0 y los 25 (0–7.4; 7.5–12.4; 12.5–17.4; 17.5–19.9; 20–22.4; 22.5–25). Estas irregularidades en parte reflejan la herencia de las escalas y en parte el constructo de la prueba, que concede distinta importancia a las diferentes dimensiones. En estas escalas los evaluadores primero hacen una estimación global de la banda a la que piensan que pertenece un candidato y luego otorgan una puntuación de la mencionada horquilla. Las diferentes puntuaciones dentro de una misma banda no influyen en la calificación final del candidato. Así pues, en el caso de la banda A1 de Competencia Pragmática, por ejemplo, una puntuación de 0 tiene el mismo peso en la calificación final del candidato que una puntuación de 10.4. Esto confiere al evaluador cierto margen de decisión y, al mismo tiempo, permite interpretar cualquier puntuación dentro de una banda como perteneciente a una misma categoría. Como decimos, a pesar de la aparente subjetividad de estas escalas,

su análisis mediante el modelo MFRM (*vid.* 2.1.2) ha demostrado que funcionan correctamente, como se demuestra en la figura 25 y la tabla 16 de más abajo, de las que se excluye la banda C2, y en cuya interpretación profundizaremos en la sección 4.2.4.2.

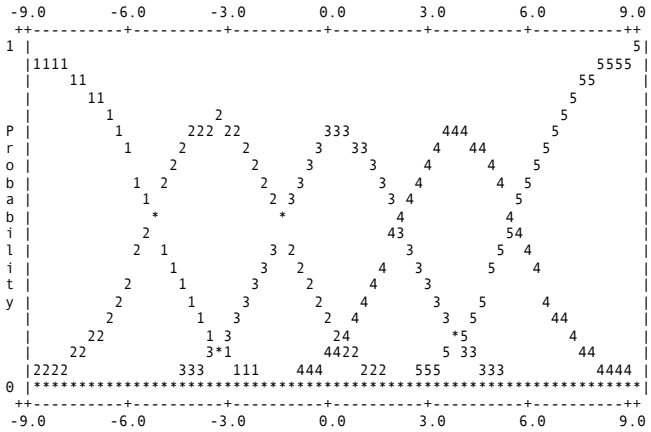


Figura 25. Curvas de probabilidad A1–C1 de la Universidad de Antioquia

DATA			QUALITY CONTROL			RASCH-ANDRICH	EXPECTATION	MOST	RASCH-	Cat	Obsd-Expd	Andrich
Category Counts		Cum.	Avge	Exp.	OUTFIT	Thresholds	Measure at	PROBABLE	THURSTONE	PEAK	Diagnostic	Thresh.
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category -0.5	from	Thresholds	Prob	Residual	Displac
1	64	48	8%	8%	-5.80	-5.65	.9			(-6.61)	low	low	100%		
2	146	146	25%	33%	-2.99	-2.95	.8	-5.53	.21	-3.56 -5.56	-5.53	-5.54	78%		-.01
3	221	221	38%	71%	.27	.28	.8	-1.64	.15	.10 -1.64	-1.64	-1.64	74%	.6	.00
4	140	140	24%	94%	2.81	2.59	1.0	1.87	.13	3.57 1.85	1.87	1.85	74%		.00
5	65	33	6%	100%	4.45	4.98	1.5	5.30	.23	(6.39) 5.38	5.30	5.32	100%		.01
										(Mean)	(Modal)	(Median)			

Tabla 16. Estadísticas de las bandas A1–C1 de la Universidad de Antioquia

4.2.2. Confección de descriptores

Un buen descriptor es un *haiku* que, como tal, está dotado de densidad semántica a pesar de su limitada extensión.

Los descriptores pueden ser desarrollados por una sola persona, por un equipo de expertos o por varios equipos que se distribuyan la confección de las distintas dimensiones y que, en cualquier caso, deben responder ante el líder del proyecto, del que ya hablamos en la sección 4.2.1. Es recomendable que, si se están desarrollando en paralelo varias escalas para una misma prueba, se tengan en cuenta aquellas dimensiones que son compartidas por distintas escalas, como hemos comentado en la sección 4.2.1.2 al hablar de un modelo de diseño esca-

lable. Ahorraremos tiempo si las dimensiones y sus respectivos descriptores se diseñan una sola vez y se adaptan a todas las escalas en las que se deban usar. En un proyecto para las escuelas oficiales de idiomas de la Comunidad Foral de Navarra, España, diseñamos seis escalas distintas para:

- Producción de textos escritos
- Coproducción de textos escritos
- Producción de textos orales
- Coproducción de textos orales
- Mediación escrita
- Mediación oral

Durante el desarrollo de las escalas se observó que, por ejemplo, las escalas de Producción de textos orales y Producción de textos escritos compartían las dimensiones de Capacidad Léxica y Capacidad Gramatical. Las escalas de Mediación escrita y Mediación oral compartían las dimensiones de Gestión del contenido y Transmisión del contenido, etc. Por lo tanto, era posible reutilizar los descriptores de unas escalas en otras con pequeñas adaptaciones. Estas coincidencias no son fortuitas y deben ser previstas al tiempo que se definen las dimensiones. En caso de trabajar con más de una escala a la vez, tener una visión de conjunto no solo favorece la consistencia interna entre las distintas escalas, sino que, como vemos, ayuda a economizar esfuerzos.

No es necesario reinventar la rueda durante la redacción de descriptores. De hecho, puede ser contraproducente. Es conveniente contar con un texto normativo que sirva de referencia. En Europa, estos documentos son el MCER (Consejo de Europa, 2001) y el MCERCV (Consejo de Europa, 2020), de los que hemos hablado en la sección 3.2. En España, por ejemplo, la práctica totalidad de los currículos de enseñanza de lenguas está diseñada a partir de la definición de los niveles establecida por dichos documentos, al igual que ocurre con los exámenes de dominio. Algunas instituciones, como por ejemplo Cambridge University Press and Assessment, que contaba con una *suite* de exámenes bien consolidada anterior al MCER (Consejo de Europa, 2001), ha hecho grandes esfuerzos por referenciar sus pruebas a los niveles A1–C2 (Cambridge University Press and Assessment, 2023), llegando incluso a adaptar sus escalas (Ffrench, 2003). Si se leen las escalas de las pruebas FCE, CAE o CPE de Cambridge Assessment, las de DELF y DALF de la Alianza Francesa o las del PLIDA de la Dante Alighieri, es fácil identificar descriptores que han sido extraídos directamente del MCER (Consejo de Europa, 2001). Dado que estos descriptores están bien diseñados y ya han sido previamente validados cualitativa y cuantitativamente (North, 2000; 2007), lo más inteligente es usarlos con las modificaciones necesarias. Pueden existir, no obstante, contextos en los que sea necesario el uso de textos normativos diferentes al MCER (Con-

sejo de Europa, 2001), como por ejemplo, América o Asia, donde este tiene una penetración desigual. Un ejemplo normativo distinto al MCER (*ibid.*) es el de los niveles de dominio utilizados por la OTAN para sus cuerpos militares (NATO Standardization Office, 2016), que los distintos países miembros de esta organización utilizan en virtud de los acuerdos STANAG.

Tanto si se parte de un texto normativo como si es necesario redactar descriptores *ex novo*, hay varias cuestiones básicas que se han de tener en cuenta en dicha redacción, fundamentalmente, el avance de los descriptores y su sintaxis. El avance de los descriptores implica que si hemos definido una dimensión con las características A, B y C, dichas características han de avanzar en complejidad a lo largo de las bandas. Por otro lado, la sintaxis de los descriptores ha de ser clara, concreta y positiva. Se deben evitar adverbios con significado impreciso del tipo «mucho», «poco» o «frecuentemente» (¿cómo se determina lo que es «mucho», «poco» o «frecuentemente»?). Se han de evitar referencias a aspectos que no queden descritos por las propias escalas o que dependan del contexto. Por ejemplo, el MCER (Consejo de Europa, 2001) y el MCERVC (Consejo de Europa, 2020) contienen descriptores que incluyen las expresiones «usa los conectores más habituales» o «comprende [...] más allá de su propio campo de especialidad». Los conectores más habituales en una lengua no tienen por qué ser los más habituales en otra, ni se suele conocer cuál es el campo de especialidad de un candidato en exámenes de dominio. Por último, la sintaxis ha de ser enunciativa afirmativa, algo que choca con el uso de bandas 0, que suelen estar enunciadas en negativo, es decir, indicando aquello que el candidato no puede hacer y no aquello que sí sabe hacer.

Durante la redacción de los descriptores se pueden utilizar hojas de cálculo que permitan apreciar la evolución de los descriptores y trazar su origen. Los descriptores suelen estar sujetos a modificaciones a lo largo de su confección, por lo que es útil desarrollar sistemas basados en colores que diferencien las distintas versiones o las modificaciones realizadas. Por ejemplo, las celdas de la primera versión de la escala pueden ser de un color, las de la segunda versión de otro, etc., o incluso se pueden utilizar colores para marcar qué partes de los descriptores se han modificado. En caso de que los descriptores estén extraídos total o parcialmente del MCER (Consejo de Europa, 2001), del MCERVC (Consejo de Europa, 2020) o de algún otro texto normativo, es recomendable acompañarlos con referencias a la tabla de la que se extrajeron o identificarlos con la página del documento del que hayan sido extraídos. Esto permitirá trazar su origen si en el futuro se desea revisar las escalas.

4.2.3. Validación cualitativa

Los descriptores se deben someter al juicio de colegas con experiencia en evaluación una vez redactados y maquetados. Esta puede ser una nueva oportunidad

para generar consenso entre los diferentes colectivos implicados. Por ejemplo, podemos intentar que esta primera versión de la escala sea revisada por un grupo representativo de los evaluadores que serán usuarios finales. De estos colegas se puede extraer información interesante mediante entrevistas.

Durante la revisión cualitativa es recomendable centrarse en cuestiones específicas tales como la redacción de los descriptores (errores sintácticos, ortográficos o ambigüedades), la adaptación de los textos normativos (¿reflejan nuestras escalas los textos de los que emanan y el constructo de la prueba?) y en cuestiones formales (por ejemplo, si la maquetación favorece un uso intuitivo y ágil de la herramienta). Cuando se pide una revisión cualitativa a colegas que no han participado en las tomas de decisiones iniciales (sobre el tipo de escala, las dimensiones o el número de bandas) es posible que haya comentarios sobre cuestiones centrales debatidas con anterioridad acerca las que ya se ha tomado una decisión. Estos comentarios han de ser considerados en un plano secundario. De lo contrario, se entrará en un bucle que nos llevará a replantearnos la estructura misma de las escalas una y otra vez, y que nos impedirá avanzar. Si hemos llevado a cabo las fases anteriores del diseño de forma correcta, podremos argumentar por qué las escalas tienen una estructura y no otra e indicar a nuestros expertos revisores sobre qué aspectos sí deseamos *feedback* y sobre cuáles no. Como ya hemos dicho, rara vez llueve a gusto de todos durante el diseño de unas escalas. Lo importante es generar un consenso colegiado y riguroso en el que, aunque haya que hacer concesiones en un sentido o en otro, las decisiones tengan una fundamentación apropiada y se mantengan en el tiempo.

Una vez recibidos estos comentarios, el líder del proyecto ha de filtrarlos e incorporar cualquier modificación que beneficie la claridad y funcionalidad de la escala. De esta manera, habrá quedado lista para la fase de validación cuantitativa.

4.2.4. Validación cuantitativa

Esta fase, de vital importancia, es tal vez la más ajena a la mayoría de evaluadores, dado que implica el uso de la estadística.

Para la validación cuantitativa de nuestras escalas usaremos el modelo MFRM. En la sección 2.1 ya hablamos de la importancia de la medición y de cómo la psicometría puede ser útil a la hora de convertir palabras en números interpretables que fundamenten nuestro argumento de validez. En la sección 2.1.2 indicábamos que el MFRM es un sistema probabilístico, es decir, un modelo estadístico que nos permite hacer proyecciones de cómo se comportará nuestro instrumento de medición (ítems o escalas) en diferentes contextos, basándonos en los datos de un sencillo pilotaje llevado a cabo con unos pocos evaluadores y candidatos. Si los datos que obtenemos mediante MFRM en esta fase son estables, podemos concluir que nuestra escala también lo es.

Existen formas altamente sofisticadas de validar cuantitativamente una escala (véase, por ejemplo, Knoch, 2009). No obstante, dicha sofisticación es poco práctica si se cuenta con recursos limitados. Existen formas más sencillas (Linacre 1999b; 2002) que son las que recomendamos utilizar. Para explicar en qué consisten estas formas de validación cuantitativa, en los párrafos que siguen hablaremos primero de cómo obtener los datos necesarios y, después, de los análisis que se han de realizar a partir de dichos datos.

4.2.4.1. Obtención de datos

Para realizar una validación cuantitativa necesitamos que distintos evaluadores usen las escalas para calificar a algunos candidatos. Es recomendable que, previamente a la calificación de los candidatos, los evaluadores participen en una sesión de familiarización con las escalas. Las muestras a calificar pueden extraerse de forma aleatoria de candidatos con el nivel evaluado y con niveles adyacentes tanto por debajo como por encima. Así pues, si estamos validando unas escalas de nivel C1 lo ideal sería hacerlo partiendo de muestras de los niveles B2, C1 y C2.

Si bien no existe consenso sobre cuál es el número ideal de evaluadores y candidatos necesarios (Anthoine *et al.* 2014:8) para este tipo de análisis, podemos obtener datos relevantes con un mínimo de dos evaluadores (A y B) y treinta candidatos (Linacre, comunicación personal). Si es posible disponer de más evaluadores y candidatos, tanto mejor. Será importante que los distintos evaluadores califiquen las mismas muestras. Es decir, el evaluador A tendrá que calificar todas las muestras de los candidatos del uno al treinta, y lo mismo ha de hacer el evaluador B. La idea es obtener dos fotografías distintas de la misma cosa. Es recomendable leer Eckes (2009) antes de lanzarse a la recogida de datos.

Finalmente, los datos obtenidos pueden ser recopilados por el líder del proyecto y volcados a un archivo de especificaciones de *Facets* (Linacre, 2024a), que debería tener una forma similar a la figura 26:

```
Title = EJEMPLO DE ESPECIFICACIONES PARA ESCALA
Facets = 3; la definición de lo que es una faceta está en el TUTORIAL 1 sección 31 y
siguientes
Inter-rater = 1
Positive = 2; ubicamos aquí las facetas orientadas positivamente
Noncentered = 2; se mide desde el centro de las otras y esta faceta flota TUTORIAL
2 punto 10
; Vertical = 1N, 2N, 3A ;
Arrange = mN; ordena el orden de las medidas de la Tabla 7
Inter-rater = 1; la faceta 1 es la del calificador
pt-biserial = yes ; point-measure correlation. Usado para analizar la consistencia de
las evaluaciones de un evaluador con el resto. Los valores menores de .30 indican
inconsistencia
gstat = yes
Models=
?,?,?,R5 ; cada ? es una faceta que interactúa con el resto para generar puntuaciones
con un valor máximo de 5

*
Labels=
1, Evaluador
1=R1; MARIA
2=R2; PETER
3=R3; ESTEBAN

*
2, Candidatos
1-30

*
3, Dimensión
1=LANGUAGE
2=DISCOURSE
3=SOCIOPRAGMATICS
4=TASK RESPONSE

*
data=
1,1,(1-4),5,4,5,3
1,2,(1-4),4,4,4,4
1,3,(1-4),3,3,2,3
[...]
2,1,(1-4),5,4,4,3
2,2,(1-4),4,3,4,3
2,3,(1-4),3,2,2,3
[...]
```

Figura 26. Ejemplo de archivo de especificaciones para *Facets*

Estos archivos de especificaciones se crean para cada análisis y, dado que su preparación no es intuitiva, el programa *Facets* (Linacre, 2024a) cuenta con tutoriales exhaustivos mediante los que aprender a componerlos. Los archivos de especificaciones contienen las características básicas de nuestras escalas y los datos a ellas vinculadas. Cualquier dato que siga a un punto y coma (;) es ignorado por el programa, lo que nos permite anotar datos que, aunque relevantes (como por ejemplo el nombre de los evaluadores), no deseamos que aparezcan en los análisis finales. Aunque hemos usado acentos en algunas de las palabras incluidas en la figura 26, lo recomendable es no usarlos porque pueden generar problemas en el análisis. El primer dato importante que tenemos en el archivo de especificaciones es el de las facetas que interactúan en el análisis, que en este caso son tres: evaluadores, candidatos y dimensiones. El concepto de facetas, aunque intuitivo en esencia, puede llevar a confusión y, por tanto, es recomendable realizar una pequeña reflexión sobre cuáles son estas antes incluso de comenzar a reunir datos, algo que el propio programa *Facets* (*ibid*.) explica en su tutorial. Definidas las facetas, tendremos que indicar al programa cómo interactúan unas con otras según el modelo, que en este caso viene descrito por la expresión ?,?,?,R5. Cada símbolo de interrogación representa a una faceta. Lo que esta expresión indica es que cualquier elemento de nuestra primera faceta (evaluadores) puede interactuar con cualquier elemento de la segunda (candidatos) y con cualquier elemento de la tercera (dimensiones) para obtener una puntuación que, como máximo, será de 5 (R5), que corresponde a la banda 5 de nuestras escalas. A continuación, indicamos al programa los candidatos y las dimensiones. La última sección relevante son las calificaciones que los evaluadores han otorgado a los candidatos. La línea 1,1,(1-4),5,4,5,3 indica que el evaluador 1 otorgó al candidato 1 en las dimensiones 1–4 las calificaciones de 5, 4, 5 y 3 respectivamente; la línea 1,2,(1-4),4,4,4,4 indica que el evaluador 1 otorgó al candidato 2 en las dimensiones 1–4 las calificaciones de 4, 4, 4 y 4, etc. Tras introducir todos los datos del evaluador 1 se introducen las calificaciones que el evaluador 2 otorgó a los mismos candidatos y, así, la línea 2,1,(1-4),5,4,4,3 indica que el evaluador 2 otorgó al candidato 1 en las dimensiones 1–4 las calificaciones de 5, 4, 4 y 3, etc. Al analizar estas especificaciones obtendremos un archivo con diferentes tablas a través del cual podremos observar cómo están funcionando nuestras escalas, de lo que hablamos a continuación.

4.2.4.2. Análisis estadísticos

Linacre (1999b) indica que, a la hora de analizar la utilidad de las bandas (o categorías) y, por extensión, nuestras escalas, debemos comprobar si:

a. tenemos al menos diez observaciones de cada banda
b. existe una distribución regular de las observaciones
c. las medias de habilidad en las bandas avanzan monótonamente
d. los valores de ajuste al modelo son menores de 2.0
e. la calibración de los peldaños avanza
f. las evaluaciones suponen medidas y viceversa
g. las dificultades de las bandas avanzan entre uno y cinco *logits*

Además de estas comprobaciones, estimamos también interesante investigar:

h. la relación entre candidatos, dimensiones y evaluadores, y
i. la fiabilidad de los evaluadores

Veamos ahora lo que cada una de estas comprobaciones implica y cómo interpretar los datos generados por *Facets* (Linacre, 2024a) que, en su mayoría encontraremos en la tabla *Category statistics* que reproducimos, para dos escalas distintas, en la tabla 16 de la sección 4.2.1.4 y en la tabla 17 de más abajo.

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	18	**18**	5%	5%	**-6.03**	**-5.93**	**.9**			(-5.94)		low	low	100%
2	47	**47**	13%	18%	**-1.55**	**-1.64**	**1.0**	**-4.86**	.41	-2.82	-4.89	-4.86	-4.88	79%
3	96	**96**	27%	45%	**.99**	**1.05**	**.8**	**-.81**	.23	.51	-.90	-.81	-.85	65%
4	101	**101**	28%	73%	**2.81**	**2.81**	**1.0**	**1.89**	.16	2.86	1.76	1.89	1.82	56%
5	98	**98**	27%	100%	**4.76**	**4.74**	**1.0**	**3.79**	.17	(4.96)	4.07	3.79	3.90	100%
										(Mean)		(Modal)	(Median)	

Tabla 17. Estadísticas de las bandas de una escala

a. Tenemos al menos diez observaciones de cada banda

Mediante esta comprobación averiguaremos si las distintas bandas se han usado el número de veces necesario y si, por tanto, hemos facilitado al programa suficiente información sobre nuestras escalas. Si una banda no se usa nunca no se puede hacer ninguna apreciación sobre ella, por lo que es necesario que cada banda sea usada al menos diez veces.

La información que necesitamos para llevar a cabo esta comprobación se encuentra en la columna *Category Counts Used* de la tabla 17. En dicha columna vemos el número de observaciones (i.e. ocasiones en las que se ha usado) de cada banda. Por ejemplo, vemos que para la banda 1 tenemos 18 observaciones; para la banda 2 tenemos 47; para la 3 tenemos 96 observaciones; 101 para la banda 4 y, finalmente, 98 para la banda 5. Así pues, el resultado tras la primera comprobación es positivo, dado que todas las bandas han sido usadas al menos diez veces

y, por lo tanto, el programa ha recibido suficiente información sobre la manera en que funcionan. Las escalas de la tabla 16 también cumplen con este requisito.

Si se ha definido una banda 0, como mencionábamos en las secciones 4.2.1.1 y 4.2.1.3, es posible que en este punto nos encontremos con un número insuficiente de observaciones que nos impida hacer inferencias fiables sobre su funcionamiento.

b. Existe una distribución regular de las observaciones

Un uso uniforme de las bandas es lo óptimo para su calibración. Linacre (1999b:110–111; 2002:7–8) indica que, en las distribuciones unimodales como la nuestra, las bandas centrales son las que se suelen usar más veces.

Así pues, en la columna *Category Counts Used* deberíamos ver que el número de observaciones es similar en todas las bandas y que los números más altos están en las bandas centrales. Que exista un solo pico de observaciones en una banda es lo que las hace unimodales. En la tabla 17, de más arriba vemos que las bandas 3, 4 y 5 cuentan con un número similar de observaciones (96, 101 y 98, respectivamente), mientras que las bandas 1 y 2 se usan con menos frecuencia (18 y 47 veces, respectivamente). La conclusión es que en la escala analizada las observaciones no son completamente regulares. Descontextualizada, esta información podría hacer sospechar que existe sesgo en la definición de las bandas. No obstante, sabemos que las muestras pertenecían a un examen de dominio B1 al que los candidatos se presentaban para conseguir el nivel mínimo que les permitiese finalizar sus estudios de grado. Dada la repercusión que la calificación de esta prueba tenía en el futuro de los candidatos, tiene sentido pensar que la mayoría se presentaban a ella estando bien preparados. Esto explicaría el uso más frecuente de las bandas más altas y el hecho de que la banda 4 (la inmediatamente superior al aprobado) sea la más usada de todas (101 veces). Que este pico de 101 observaciones aparezca en las bandas centrales es lo deseable según Linacre (1999b:110; 2002:7).

En la tabla 16 de la sección 4.2.1.4 encontramos una distribución más regular. Vemos que hay 48 observaciones usadas para la banda 1 (A1), 146 para la banda 2 (A2), 221 para la banda 3 (B1), 140 para la banda 4 (B2) y 33 para la banda 5 (C1). Esta distribución es, además, claramente unimodal con un pico de 221 observaciones bien diferenciado en la banda 3 (B1), por lo que estas escalas también cumplen con los requisitos relativos a la distribución de las observaciones.

c. Las medias de habilidad en las bandas avanzan monótonamente

En el diseño de cualquier escala está implícito que los candidatos que ubicamos en bandas inferiores tienen menos habilidad que los candidatos de bandas supe-

riores. La asunción teórica es que los candidatos que ubiquemos en la banda 1 tendrán una puntuación media inferior a la de los candidatos de la banda 2, que a su vez tendrán una puntuación media global inferior a la de los que estén en la banda 3, etc. Estas medias de habilidad son un indicador empírico del contexto en que cada banda se usa (Linacre, 1999b:111; 2002:8). Decimos que nuestra escala muestra avance monótono cuando la media de habilidad de los candidatos de una banda es inferior a la media de habilidad de los candidatos de la banda superior. De no ser así, no sabríamos qué significa que un candidato obtenga una puntuación superior a la de otro.

Para comprobar si nuestras bandas avanzan monótonamente debemos fijarnos en la columna *Avge Meas* (abreviación de *average measure*, media de habilidad) de la tabla 17. La columna *Avge Meas* nos dice que los candidatos a los que se ubicó en la banda 1 obtuvieron una media de puntuación de -6.03, que es una media inferior a la de -1.55 que obtuvieron los candidatos de la banda 2 y que a su vez es inferior a la de 0.99 de los candidatos de la banda 3, etc. Si seguimos haciendo la comprobación en el resto de las bandas, vemos que la media de los candidatos en todas ellas es superior cuanto mayor es la dificultad de la banda, por lo que podemos decir que, efectivamente, nuestras bandas avanzan de forma monótona. Lo mismo ocurre en la columna *Avge Meas* de la tabla 16.

En la tabla 17 observamos que el avance es irregular y que el salto que hay entre -6.03 y -1.55 es mucho mayor (casi el doble) que el que hay entre -1.55 y 0.99. Algo similar ocurre en la tabla 16 donde se pasa de -5.53 a 1.64. Según Linacre (1999b:112; 2002:9), esto puede deberse a problemas con el uso de la escala o puede no ser más que un reflejo de la relación entre la dimensión concreta y la distribución de las muestras. En cualquier caso, si se sospecha que pudiera deberse a un mal uso de las escalas es recomendable revisar tanto la fiabilidad de los evaluadores (de la que hablaremos más adelante) como la redacción de los descriptores.

d. Los valores de ajuste al modelo son menores de 2.0

El modelo de probabilidad MFRM usado por *Facets* (Linacre, 2024a) asume determinado nivel de aleatoriedad en las mediciones obtenidas y compensa que pueda haber mediciones que no respondan a determinados patrones esperados. Todos los datos que suministramos al modelo aportan información, y cuando dichos datos son excesivamente aleatorios (muy impredecibles) o muy poco aleatorios (muy predecibles), se genera lo que se conoce como «ruido» en las observaciones, un ruido que, si es excesivo, puede llegar a impedir una correcta interpretación de los resultados.

Facets (Linacre, 2024a) dispone de un control de calidad para niveles excesivos de ruido. Como quedó descrito en la sección 2.1.2.2, este control de calidad son los valores de ajuste de nuestras medidas, que son sensibles a los patrones de respuestas improbables. Según Linacre (1999b:112–113; 2002:9–10), este detector de ruido debe estar siempre por debajo de 2 y tender a 1. Cuanto más cerca estén estos valores de 1, menos ruido habrá detectado el modelo en nuestras observaciones y, por lo tanto, más seguridad podremos tener de que las inferencias hechas a partir de nuestros datos son correctas. Cuanto más nos alejemos de 1, más ruido habrá detectado el modelo y menos seguros podremos estar de nuestras inferencias.

Al observar la tabla 17, en la columna *Outfit Mnsq* vemos que los valores oscilan entre 0.8 (el más bajo) y 1 (el más habitual), lo cual es buena señal. Estas cifras demuestran que los datos que hemos suministrado al modelo matemático no han generado apenas ruido y no están causando interferencias que distorsionen nuestra interpretación de los datos. Han aportado información valiosa y, por tanto, podemos tener la certeza de que las inferencias hechas a partir de nuestro análisis serán fiables. En la tabla 16 observamos que la mayoría de los valores son óptimos o muy buenos, con la excepción del 1.5 correspondiente a la banda 5 (C1) que, con todo, sigue dentro de los parámetros esperados.

e. La calibración de los peldaños avanza

Para entender esta y algunas de las comprobaciones que siguen será útil volver a imaginar que nuestra herramienta de medida es una escalera y que sus bandas son peldaños. La banda 1 sería el peldaño más bajo y la banda 5 el más alto, y los candidatos a nuestras pruebas podrán subir tantos peldaños como su habilidad les permita. En lugar de «peldaño», Linacre (1999b:114–115; 2002:10–11) usa la palabra «paso» (*step*) que toma prestada de Wright y Masters (1982:40–41) (*vid*. Linacre, 1999a).

Wright y Masters (1982:40–41) hablan de peldaños (*steps* en el original) para distinguir entre puntuaciones dicotómicas (correcto vs. incorrecto) y otras en las que se pueden establecer diferentes niveles de habilidad que subsumen al anterior. Esto es precisamente lo que ocurre en cualquier escala, en la que un candidato solo puede ser ubicado en la banda 2 si ha superado el nivel de competencia descrito en la banda 1, o en la banda 3 si ha superado el nivel de competencia descrito en la banda 2, etc. En esta lógica, cuanto mayor sea la habilidad de un candidato, más probable ha de ser que se le califique con una banda más alta (Linacre, 1999b:114; 2002:10) o, dicho de otra forma, más peldaños de la escalera podrá subir. Veamos cómo se manifiesta esto en las curvas de probabilidad de las distintas bandas (*vid*. 2.1.2).

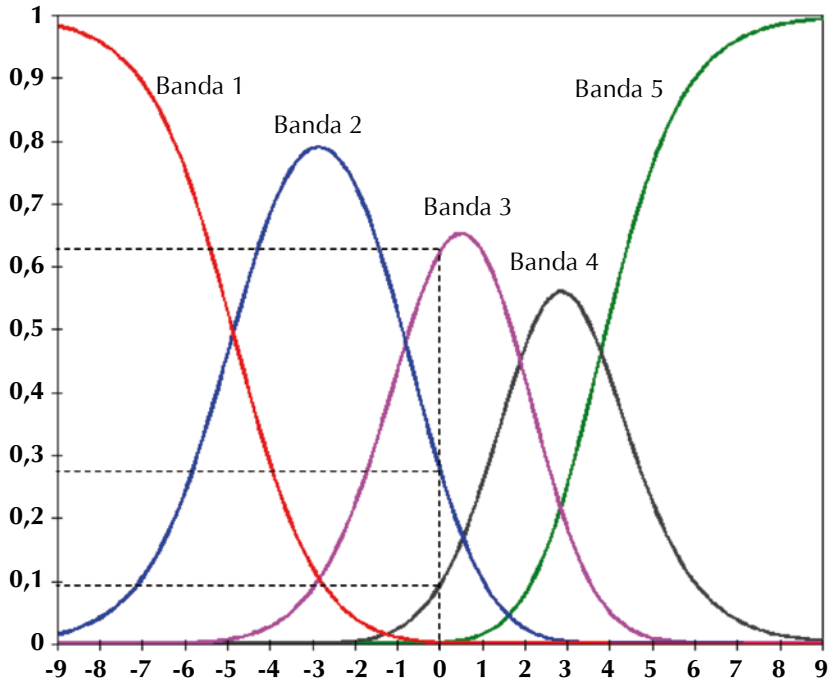


Figura 27. Curvas de probabilidad de las bandas de una escala

Las curvas de la figura 27 están construidas a partir de un eje vertical, que es el que indica qué probabilidad hay de que a un candidato se le califique con una banda, y un eje horizontal, que es el que refleja las posibles habilidades de los candidatos. Según la figura 27, por ejemplo, un candidato con una habilidad de - 9 en el eje horizontal tendría una probabilidad de algo más del 0 % de ser calificado con la banda 2 y, al mismo tiempo, tendría una probabilidad de casi el 100 % de ser calificado con la banda 1. De igual forma, un candidato con una habilidad de 0 en el eje horizontal tendría una probabilidad algo menor del 10 % de ser calificado con la banda 4, una probabilidad de aproximadamente el 28 % de ser calificado con la banda 2 y una probabilidad aproximada del 63 % de ser calificado con la banda 3. Estos porcentajes se obtienen trazando una línea vertical desde el punto 0 del eje horizontal (representado por la línea de puntos en la figura 27) y proyectando al eje vertical de probabilidad los puntos de corte entre dicha línea y las bandas. Se pueden trazar cualesquiera otras líneas verticales para averiguar las probabilidades de candidatos con otros niveles de habilidad.

Nuestra escala ha de ser una escalera que avance de manera continua, no una escalera que, digamos, tenga dos peldaños de subida, uno de bajada y otros dos de subida. Dicha continuidad es la que analizamos con esta comprobación. Lo

que observamos en la figura 27 es que cuanto mayor es la habilidad de los candidatos, más probable es que se les califique con una banda superior. Al mismo tiempo, desde la perspectiva de las bandas observamos que todas ellas son las de más probable elección en algún punto, algo que gráficamente se aprecia en el hecho de que todas tienen un pico que sobresale por encima de los puntos de corte con las bandas adyacentes.

El desorden de los peldaños de una escala puede deberse a distintos motivos (Linacre, 1999a). Por ejemplo, puede deberse a que una banda concreta describa un segmento de la variable latente excesivamente pequeño (Linacre, 1999b:114– 115; 2002:10–11). La figura 28 (Cantó-Cerdán, 2021) muestra una escala con los peldaños desordenados, en la que la banda 2 no es modal, es decir, no es la de más probable elección en ningún momento o, como decíamos más arriba, no tiene un pico que sobresalga por encima de la línea de corte con las bandas adyacentes.

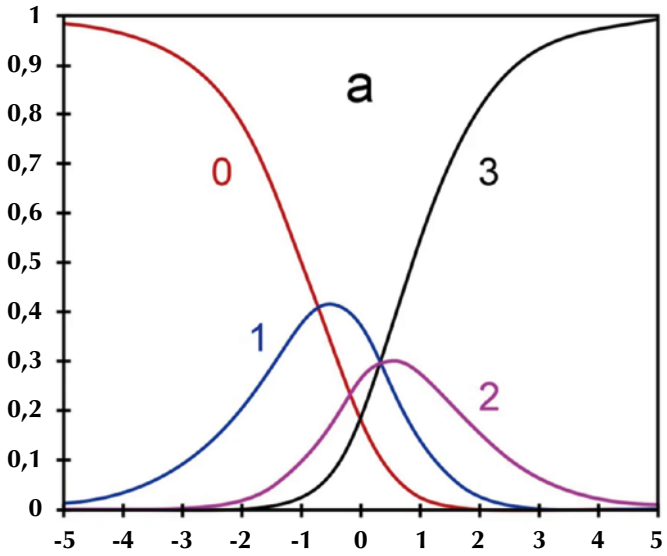


Figura 28. Peldaños desordenados

La calibración de los peldaños está, pues, íntimamente ligada a los puntos en los que las curvas de probabilidad de nuestras bandas se cruzan. Estos puntos de corte indican los niveles de habilidad en los que un candidato tiene la misma probabilidad de ser calificado con cualquiera de las dos bandas adyacentes, se conocen como umbrales Rasch-Andrich (Linacre, 1999a), y se espera que siempre aparezcan vinculados a niveles ascendentes de habilidad. Se podría decir que estos puntos de corte entre bandas son los que separan un peldaño de otro. Así, en la figura 27 vemos que la intersección entre la banda 1 y la banda 2 está clara-

mente separada de la intersección entre la banda 2 y la banda 3, mientras que en la figura 28, sin embargo, la intersección entre la banda 1 y la 2 está en el mismo lugar que la intersección entre la banda 2 y la banda 3. Incluso pueden darse casos en los que la intersección entre bandas que supuestamente reflejan más habilidad esté más a la izquierda del eje horizontal que la intersección de bandas que teóricamente reflejan menos habilidad. *Facets* (Linacre, 2024a) especifica los puntos exactos en los que las curvas de las bandas se cortan. Como se observa en la columna *RASCH-ANDRICH Thresholds Measure* de las tablas 16 y 17, todos los valores (i.e. puntos de corte entre bandas adyacentes) avanzan monótonamente, es decir, cualquiera de ellos es mayor que el que le antecede, justo lo que deseamos encontrar en nuestras escalas.

Así pues, dado que los distintos umbrales Rasch-Andrich de nuestra escala se encuentran separados entre sí y, por tanto, las distintas curvas de probabilidad de las bandas dibujan «montañas» con un pico que sobresale sobre los puntos de corte de las bandas adyacentes, consideramos que los peldaños de nuestras escaleras se encuentran ordenados y que avanzan unos con respecto a otros.

f. Las observaciones suponen medidas y viceversa

Este aspecto es particularmente útil para confirmar la validez del constructo de nuestra escala (Linacre, 1997). En general, los usuarios de nuestras escalas suponen que cualquier observación hecha a través de ellas equivale a una medida subyacente y viceversa (Linacre, 1999b:11).

Podremos decir que eso es cierto si en las tablas 16 y 17 las medidas de la columna *Avge Meas (average measure*, habilidad media) son cercanas a las de *Exp Meas (expected measure*, habilidad esperada). Hasta donde sabemos, no existe un valor máximo o mínimo de referencia en la literatura (al contrario de lo que ocurría con los valores de ajuste) más allá de lo descrito por Linacre (1997), que considera relevante una diferencia de 0.63 *logits*. Por lo tanto, este factor está relativamente abierto a la interpretación del equipo de confección de escalas. En el caso de las tablas 16 y 17 vemos que los valores medios obtenidos y los esperados están muy cerca unos de otros, por lo que podemos interpretar que, en general, las observaciones hechas con ambas escalas implican medidas y viceversa.

g. Las dificultades de los peldaños avanzan entre 1 y 5 *logits*

Antes hemos visto que una de las propiedades de nuestra escala ha de ser que sus peldaños avancen. Ahora vamos a comprobar la altura de dichos peldaños. Vamos a comprobar si la altura que separa al primer peldaño del segundo es la misma que separa al segundo del tercero y así sucesivamente. La separación necesaria entre peldaños depende de cuántos contenga nuestra escala. Así, por ejemplo, en una escala de tres bandas, la distancia mínima entre peldaños ha de ser de 1,4 *logits*, mientras que en una escala con cinco bandas ha de ser de 1 *logit*. Por otro

lado, la distancia máxima entre dichos peldaños ha de ser de 5 *logits* (Linacre, 1999b:117–20; 2002:12–14) independientemente del número de estos que haya.

Si nuestros peldaños no tienen la distancia mínima de 1 *logit* puede que hayamos diseñado bandas que describan un segmento muy pequeño de la variable latente. En estos casos puede interesar unir en una sola aquellas bandas que estén poco separadas. Por otro lado, si diseñamos bandas que tengan una distancia excesiva entre sus puntos de corte podríamos dar lugar a bandas que describen un segmento de la variable latente excesivamente extenso, con «ángulos muertos» que podrían llevar a una pérdida de precisión en las mediciones (Linacre, 1999b:119; 2002:12–13).

Para comprobar si las distancias mínimas y máximas son las adecuadas, debemos acudir de nuevo al concepto de los umbrales Rasch-Andrich y a la columna *RASCH-ANDRICH Thresholds Measure*. En este caso nos concentraremos en los valores descritos en la tabla 17. En la columna *RASCH-ANDRICH Thresholds Measure* de la tabla 17 vemos que el primer punto de corte está en -4.86, el segundo en -0.81, el tercero en 1.89 y el cuarto en 3.79. Si proyectamos los puntos de corte entre las distintas bandas al eje horizontal de la figura 29, veremos que dichas proyecciones cortan el eje X (que corresponde a la habilidad de los candidatos) justo en las cifras que nos ha dado la tabla 17. Es decir, la figura 29 es una representación gráfica de los valores de la columna *RASCH-ANDRICH Thresholds Measure*. Si calculamos la distancia entre dichos puntos de corte estaremos calculando la altura de nuestros peldaños y así podremos comprobar si estos están entre 1 y 5 logits.

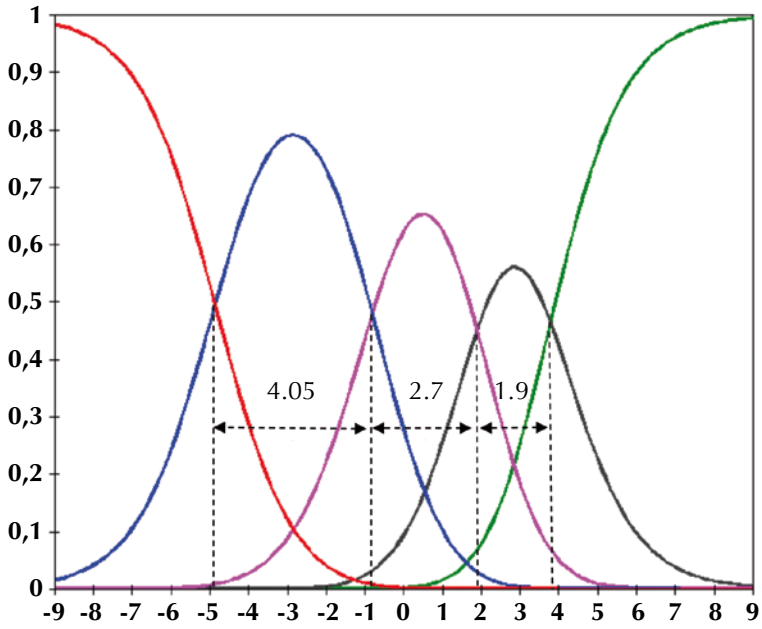


Figura 29. Separación de los umbrales Rasch-Andrich

Efectivamente, vemos que todos nuestros peldaños se separan el uno de otro dentro de los mínimos y máximos deseados. El primer peldaño tiene una altura de 4.05 *logits*, el segundo de 2.7 y el tercero de 1.9. No se pueden establecer estas medidas en los extremos porque tanto la primera como la última banda son asintóticas, es decir, tienden al infinito. Este cálculo evidencia que, aunque la separación está dentro de los valores mínimos y máximos, la distancia del segundo peldaño (4.05) representa un segmento de la variable latente mucho mayor (más del doble) que el que representa el cuarto escalón (1.9). La experiencia nos dice que es prácticamente imposible que el tamaño de los peldaños sea exactamente el mismo. No obstante, puede ser interesante revisar los descriptores de las escalas y su uso si se dan casos de diferencias notables, aun cuando estas estén dentro de los valores de referencia. Cuando dos de estos valores estén excesivamente juntos, tal vez se pueda considerar la opción de unir dos bandas en una y, cuando estén excesivamente separados, tal vez se pueda considerar la creación de una banda intermedia.

h. Relación entre candidatos, dimensiones y evaluadores

Una inspección visual de las relaciones entre candidatos, categorías y dimensiones es interesante al tiempo que una forma gráfica de presentar los resultados del análisis a aquellos agentes implicados en las pruebas que no tengan conocimientos de estadística. Esta inspección se puede realizar a través del *vertical ruler* de *Facets* (Linacre, 2024a), un mapa similar al de la variable del que hablamos en la sección 2.1.2.3 pero que, en lugar de ítems y candidatos, incluye (de izquierda a derecha) a los evaluadores, a los candidatos, las dimensiones y las bandas de nuestra escala. Tenemos un ejemplo en la figura 30.

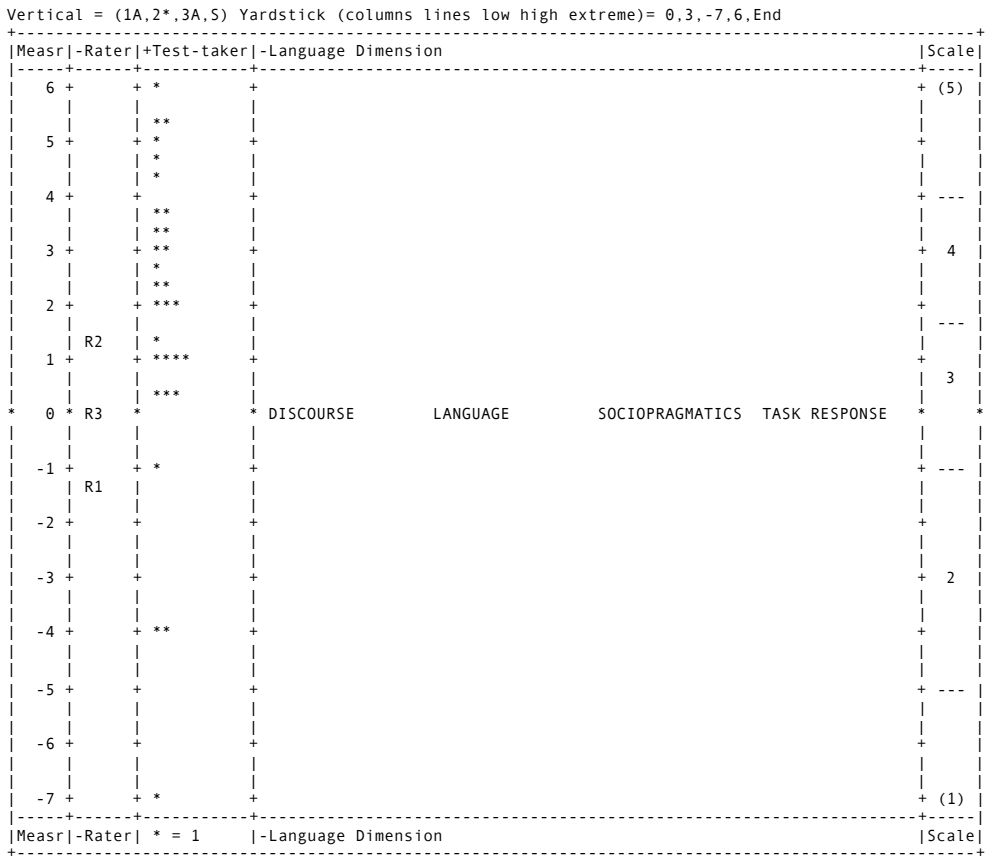


Figura 30. *Vertical ruler*

Este mapa nos ofrece información fácil de entender a simple vista. En primer lugar, por ejemplo, en la columna *Rater* se ordena a los distintos evaluadores en función de su severidad a la hora de corregir. En este caso particular, observamos que el evaluador R2 ha sido el más severo, mientras que el R1 ha sido el más benevolente. También podemos ver la disposición de los candidatos en la columna *Test-taker*. En esta columna apreciamos que la mayoría de candidatos se sitúa en la mitad superior de la columna, lo que indica que en su mayoría obtuvieron calificaciones altas, y entronca con lo comentado en el apartado «b» sobre la concentración de observaciones en bandas altas debido a la buena preparación de los candidatos. Por último, también vemos que las distintas dimensiones están ubicadas en el mismo nivel de dificultad, lo cual es una propiedad deseable en nuestras escalas, dado que no es deseable que una dimensión sea más difícil de superar que el resto.

Fiabilidad de los evaluadores

En este último apartado no estamos observando una propiedad intrínseca de la escala, sino la forma en la que los evaluadores se comportan al utilizarla. A la hora de «cuantificar el grado en el que las evaluaciones de cada evaluador son consistentes con las del resto de evaluadores [...], los valores inferiores a 0.30 permiten identificar a los evaluadores inconsistentes, en los que la ordenación de las personas difiere de la del resto de los calificadores» (Prieto, 2011:234). Para obtener estos valores hemos debido incluir en el archivo de especificaciones (figura 26) la orden *pt-biserial = yes*. Esto nos permitirá obtener la tabla 18 de más abajo. Al observar la columna *Corr PtBis* vemos que, efectivamente, los tres evaluadores han sido consistentes a la hora de ordenar a los candidatos de una forma concreta, puesto que las cifras correspondientes a cada uno (0.56; 0.57 y 0.58) son mayores de 0.30.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	**Corr. PtBis**	Exact Agree. Obs %	Exact Agree. Exp %	N	Rater
370	120	3.08	2.95	1.49	.15	.83	-1.4	.85	-1.0	1.14	**.56**	29.6	33.0	2	R2
435	120	3.63	3.56	-.04	.16	1.19	1.3	1.12	.8	.83	**.57**	36.3	43.2	3	R3
489	120	4.07	4.19	-1.44	.17	.93	-.4	.80	-.9	1.10	**.58**	32.5	36.4	1	R1
431.3	120.0	3.59	3.56	.00	.16	.98	-.2	.92	-.4		.57			Mean (Count: 3)	
48.7	.0	.41	.51	1.20	.01	.15	1.2	.14	.9		.01			S.D. (Population)	
59.6	.0	.50	.62	1.47	.01	.19	1.4	.17	1.1		.01			S.D. (Sample)	

Model, Populn: RMSE .16 Adj (True) S.D. 1.19 Separation 7.43 Strata 10.24 Reliability (not inter-rater) .98
Model, Sample: RMSE .16 Adj (True) S.D. 1.46 Separation 9.13 Strata 12.51 Reliability (not inter-rater) .99
Model, Fixed (all same) chi-squared: 165.9 d.f.: 2 significance (probability): .00
Model, Random (normal) chi-squared: 2.0 d.f.: 1 significance (probability): .16
Inter-Rater agreement opportunities: 360 Exact agreements: 118 = 32.8 % Expected: 135.0 = 37.5 %

Tabla 18. Correlación calificador-resto de calificadores

Al igual que ocurría con nuestras pruebas en su conjunto, la validación de una escala se realiza reuniendo pruebas de diferente índole que garanticen que las inferencias que hacemos de las calificaciones obtenidas mediante ellas son fiables. Recoger información de las diferentes fases de diseño (desde la sección 4.2.1 a la sección 4.2.6) es importante pero, sin duda, la información derivada del análisis estadístico de nuestras escalas es fundamental, dado que es una base sólida, objetiva y científica para demostrar que la herramienta de medida que hemos creado funciona como se espera.

Concluida esta fase de validación, tendremos información suficiente para decidir si las escalas necesitan algún ajuste, o si están listas para ser llevadas a un examen real.

4.2.5. Implementación

Tras haber comprobado que nuestras escalas son fiables y antes de que se usen en una prueba real, es importante darlas a conocer entre los candidatos y entre los evaluadores.

En ocasiones, unas escalas vienen a sustituir a otras a las que, tal vez, los candidatos a un examen ya estaban acostumbrados. Cualquier candidato debería consultar antes del examen las escalas que se usarán para calificarle. Por eso es importante difundirlas en las plataformas que habitualmente consultan los candidatos e incluso, en el caso de exámenes de mucha repercusión, puede ser interesante organizar talleres y conferencias en las que se pueda explicar cuál ha sido el proceso de diseño y validación. Este tipo de prácticas redunda en beneficio de la transparencia y refuerza la solidez de cualquier examen de domino. Puede incluso llegar a ser recomendable comenzar estas campañas de difusión meses o años antes de que las escalas sean implementadas por primera vez, de manera que los candidatos sean conocedores de los cambios con suficiente antelación. Como candidatos, este tipo de prácticas serían las que esperaríamos de un examen serio que pueda tener un impacto profundo en nuestra vida (el acceso a un permiso de trabajo o residencia, la obtención de un puesto de trabajo, etc.).

Igualmente necesario será dar a conocer las escalas a los evaluadores. La profesionalidad llevará a la mayoría de estos evaluadores a analizar de forma crítica las escalas, a identificar los cambios con respecto a versiones anteriores y a interiorizar estos cambios. No obstante, la adopción de una nueva escala ha de estar guiada por la institución responsable de las pruebas, y orientada no solo a dar a conocer los cambios, sino también a verificar que los evaluadores son capaces de utilizar de forma fiable la nueva herramienta de medida. Por ello, en este punto tendrán especial importancia los procesos de estandarización, también conocidos en inglés como *benchmarking*.

El objetivo de la estandarización es aquilatar los criterios de los evaluadores que usarán las escalas. En comunidades pequeñas se puede reunir a todos los evaluadores para trabajar en grupo. En comunidades más grandes habrá que realizar sesiones de estandarización separadas. Puede ser muy útil contar con un equipo de estandarización encargado de repetir la misma sesión con los mismos criterios y las mismas muestras, pero con distintos evaluadores. Por regla general, estas sesiones de estandarización suelen comenzar con una presentación de las escalas, que los participantes deben conocer y haber estudiado de antemano. En esta primera fase de familiarización del *benchmarking* se pueden explicitar los cambios operados en las escalas y comparar estas con versiones anteriores. Después se suele elegir una muestra cuya evaluación no presente complejidad para que los participantes la analicen y califiquen individualmente. Una vez que todos los participantes hayan calificado esta muestra se deben compartir las calificaciones de manera que se hagan evidentes aquellos puntos en los que pudiera existir disensión. En estos primeros compases puede haber distancia entre las calificaciones de unos y otros evaluadores. Lo importante en este punto es que las calificaciones queden anotadas antes de que estas sean compartidas para evitar que los evaluadores se vean influidos por lo que señalen otros compañeros y modifiquen

así su calificación original. Cuando se está inseguro sobre una calificación es fácil dejarse arrastrar por las opiniones de otros compañeros que puedan ser más o menos similares entre sí. Se deben exponer entonces las calificaciones y abrir un debate guiado para analizar dónde pueden sobrevenir los principales problemas durante las evaluaciones.

Es de vital importancia recordar con frecuencia a los evaluadores que el objeto de estos debates es el de que afinen sus puntuaciones con respecto a las escalas, no el de cuestionarlas o el de imponer una opinión sobre el resto.

Después del análisis de una primera muestra más o menos clara se puede abordar el análisis de otras muestras cuya calificación presente más dudas. El Consejo de Europa (2009:35–56) ofrece una serie de recomendaciones muy útiles para estos procedimientos que abarcan desde la preparación de las muestras necesarias hasta la estructura de los debates y el análisis de los datos obtenidos en las sesiones. La fiabilidad y consistencia de las calificaciones obtenidas en estas sesiones de *benchmarking* pueden analizarse fácilmente mediante el procedimiento descrito en la sección 4.2.4.2/«i».

4.2.6. Revisión

Se debe establecer un ciclo de revisión de las escalas (cada dos años, cada cinco, etc.) que permita responder de manera eficiente a los retos que emanen de su uso en situaciones reales. Un ciclo de revisión no implica una modificación total de las escalas, ni siquiera necesariamente una modificación parcial. En exámenes de alto impacto, cualquier modificación tiene ramificaciones que es importante considerar antes de adoptar decisiones de calado. Además, por pequeñas que sean, estas modificaciones pueden llegar a requerir una repetición de la validación cualitativa descrita en la sección 4.2.4.

Recoger los datos generados en diferentes convocatorias es una buena opción para comparar cómo se comportan los evaluadores y las escalas en todas ellas. Estos datos también nos indicarán si existe algún problema recurrente que hubiera pasado desapercibido durante la revisión cualitativa o cuantitativa, o que hubiera podido surgir como consecuencia de cambios en el contexto de la prueba (posibles cambios demográficos, modificaciones en el diseño de las tareas, interiorización insuficiente de las escalas por parte de los evaluadores o los candidatos, etc.). En caso de que fuese necesario, se deben realizar nuevas campañas de difusión y formación sobre las escalas.

EPÍLOGO

Concluye aquí una obra cuyo sentido medular ha sido el de reflexionar sobre las pruebas de certificación de dominio de lengua.

En el primer capítulo hemos intentado ofrecer un resumen de la forma en que se ha estudiado el fenómeno del habla humana desde la antigüedad hasta nuestros días. La historia nos muestra que solo en fechas recientes hemos empezado a atisbar la verdadera naturaleza del lenguaje gracias, por un lado, a que hemos vuelto nuestra mirada hacia el organismo que lo hace posible, el ser humano, y, por otro, al exponencial avance de la tecnología. Sin duda, el próximo paso en este camino de descubrimiento supondrá reflexionar sobre la inteligencia artificial y el resto de tecnologías que hoy continúan la Revolución Digital iniciada a mediados del siglo XX, cuya importancia ya iguala a la de la Revolución Industrial del XIX.

En el segundo capítulo hemos enumerado las herramientas básicas con las que se pueden transformar las palabras en números para redactar pruebas más fiables y justas. Animamos a los agentes implicados, principalmente a gestores y redactores, a adentrarse en el ajeno pero necesario territorio de la psicometría. Los animamos a avanzar en el camino hacia fórmulas más científicas, fiables e innovadoras, aunque sea partiendo del estudio indirecto del lenguaje.

Hemos volcado toda nuestra experiencia en el tercer capítulo para ofrecer soluciones prácticas y realistas al diseño de pruebas de dominio de lengua. Este capítulo aporta definiciones y ejemplos de elementos críticos (validez, constructo, especificaciones, componentes, etc.) y al mismo tiempo señala el camino a aquellos que deseen profundizar en cualquiera de dichos aspectos. La tabla 10 y la figura 19 serán particularmente útiles para el lector con una orientación eminentemente práctica.

Por último, en el capítulo cuarto hemos dedicado a las escalas de evaluación un espacio que rara vez suelen tener en publicaciones de este tipo. Proponemos en la tabla 13 una guía realista para quienes deseen crear, adaptar o validar sus escalas de forma objetiva.

Desearíamos concluir con una reflexión sobre la forma en que, a veces, los lingüistas, docentes y redactores nos refugiamos en el humanismo para huir del resto de ciencias naturales. Humanismo y ciencias naturales no son, ni mucho menos, conceptos excluyentes. De hecho, será la interacción entre ambos la que moldeará el futuro de nuestra disciplina. El humanismo es un valor transversal, no es de letras ni de ciencias. Las ciencias naturales, por su parte, son el punto de apoyo que los lingüistas necesitamos para mover el mundo. El lenguaje es el vehículo mediante el cual exploramos el cosmos, compartimos descubrimientos y cuestionamos nuestra existencia, de ahí el papel fundamental de los lingüistas en el futuro de la especie humana. En algún lugar, algo excepcional está esperando ser descubierto por un lingüista.

REFERENCIAS BIBLIOGRÁFICAS

Abad, F *et al*. (2001). Analysis of the optimum number alternatives from the Item Response Theory, *Psicothema*, 13 (1), 152–158.

Abad, F *et al*. (2011). *Medición en ciencias sociales y de la salud*, Madrid: Síntesis.

Alderson, C (2000). *Assessing Reading*, Cambridge: Cambridge University Press.

Andrich, D (1978). A rating formulation for ordered response categories, *Psychometrika*, 43, 561–573.

Allen, M y Yen, W (1979). *Introduction to measurement theory*, Monterey (CA): Brooks-Cole.

ALTE (Association of Language Testers in Europe) (2020). *Principles of Good Practice 2020*, documento digital obtenido de https://pt.alte.org/resources/Documents/ALTE%20Principles%20of%20Good%20Practice%20Online%20version%20Proof%204.pdf

Alvermann, D *et al*. (Eds) (2013). *Theoretical Models and Processes of Reading*, Newark (DE): International Reading Association.

Anthoine, E *et al*. (2014). Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures, *Health and Quality of Life Outcomes* 12 (176), 1–10.

APA (American Psychological Association) (1952). Technical recommendations for psychological tests and diagnostic techniques: preliminary proposal, *American Psychologist* 7 (8), 461–475.

Arellano, F (1979). *Historia de la lingüística. Tomo I. Desde sus orígenes hasta el siglo XIX incluido*, Caracas: Universidad Católica Andrés Bello.

Assessment Systems Corporation (2014). *XCalibre* (versión 4.2), programa informático, Stillwater (MN): https://assess.com/xcalibre

Atzlesberger, U *et al*. (2015). *Framework for the Upper Secondary Level Oral Leaving Examination at Colleges for Higher Vocational Education*, documento

digital obtenido de https://www.cebs.at/wp-content/uploads/2019/05/Austrian_framework_plurilingual_oral_exams-Druckausgabequalit-1.pdf

Audacity Team (2024). *Audacity* (versión 3.5.1), programa informático, obtenido en https://www.audacityteam.org

Bachman, L (1991). *Fundamental Considerations in Language Testing*, Hong Kong: Oxford University Press.

Bachman, L y Palmer, A (1996). *Language Testing in Practice*, Oxford: Oxford University Press.

Baddeley, A (1988). *Working Memory*, Oxford: Oxford University Press.

Baddeley, A (2000a). The episodic buffer: a new component in working memory?, *Trends in Cognitive Science* 4 (11), 417–423.

Baddeley, A (2000b). Working memory, *Current Biology* 20 (4), R136–R140.

Baddeley, A. (2003). Working memory and language: an overview, *Journal of Communication Disorders* 36, 189–208.

Bax, S (2013). The cognitive processing of candidates during reading tests: evidence from eye-tracking, *Language Testing* 30 (4), 441–465.

Bazerman, C *et al*. (2017). Taking the long view on writing development, *Research in the Teaching of English* 51, 351–360.

Benítez, B (2016). *Genes y lenguaje: aspectos ontogenéticos, filogenéticos y cognitivos*, Barcelona: Editorial Reverté.

Bennet, A (2015). *Kendo: Culture of the Sword*, Oakland (CA): University of California Press.

Binet, A, y Simon, T (1948). The development of the Binet-Simon Scale, 1905-1908, en Dennis, W (Ed) *Readings in the history of psychology*, Nueva York (NY): Appleton-Century-Crofts, 412–424.

Boeckx, C y Piattelli-Palmarini, M (2005). Language as a natural object, *The Linguistic Review* 22, 447–466.

Bolaño, R (2004). *2666*. Barcelona: Anagrama.

Bond, T y Fox, C (2007). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*, Mahwah (NJ): Lawrence Erlbaum Associates.

Brunfaut, T (2016). Assessing listening, en Tsagari, D *et al*. (Eds) *Handbook of Second Language Assessment*, Boston (MA): De Gruyter Mouton, 97–112.

Brunfaut, T (2022). Eye-Tracking as a Research Method in Language Testing, en Mohebbi, H y Coombe, C (Eds) *Research Questions in Language Education and Applied Linguistics*, Springer Texts in Education, 737–742.

Buchanan, R y Finch, S (2005). History of psychometrics, en Everitt, B y Howell, D (Eds) *Encyclopedia of Statistics in Behavioral Science*, Nueva Jersey (NJ): John Wiley & Sons, 875–878.

Buscher, G *et al*. (2010). Eye tracking analysis of preferred reading regions on the screen, *Proceedings of the 28th International Conference on Human Factors in*

Computing Systems, CHI (Computer Human Interaction) 2010, Extended Abstracts Volume, 3307–3312.

Cambridge University Press and Assessment (2023). *The Cambridge English Scale Explained*, documento digital obtenido de https://www.cambridgeenglish.org/images/210434-converting-practice-test-scores-to-cambridge-english-scale-scores.pdf

Cambridge University Press and Assessment (2024a). *Information for candidates B2 First*, documento digital obtenido de https://www.cambridgeenglish.org/Images/610341-b2-first-information-for-candidates.pdf

Cambridge University Press and Assessment (2024b). *Information for candidates C1 First*, documento digital obtenido de https://www.cambridgeenglish.org/Images/610342-c1-advanced-information-for-candidates.pdf

Cambridge University Press and Assessment (2024c). *Handbook for teachers 2023*, documento digital obtenido de https://www.cambridgeenglish.org/Images/168194-c2-proficiency-teachers-handbook.pdf

Cambridge University Press and Assessment (2024d). *Information for candidates C2 Proficiency*, documento digital obtenido de https://www.cambridgeenglish.org/Images/610343-c2-proficiency-information-for-candidates.pdf

Campbell, N y Jeffreys, H (1938). Symposium: Measurement and its importance for philosophy, *Proceedings of the Aristotelian Society, Supplementary Volumes*, 17 (1), 121–150.

Campbell, D y Fiske, D (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin* 56 (2), 81–105.

Cantó-Cerdán, M *et al*. (2021). Rasch analysis for development and reduction of Symptom Questionnaire for Visual Dysfunctions (SQVD), *Scientific Reports*, 11, 14855.

Card, N *et al*. (2024). An accurate and rapidly calibrating speech neuroprothesis, *The New England Journal of Medicine* 391 (7), 609–618.

Chalhoub-Deville, M y O'Sullivan, B (2020). *Validity*, Sheffield: Equinox.

Chomsky, N (2000). *New Horizons in the Study of Language and Mind*, Cambridge (MA): Cambridge University Press.

Chou, Y y Wang, W (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals, *Educational and Psychological Measurement* 70 (5), 717–731.

Clark, J (1979). Direct vs. semi-direct tests of speaking ability, en Briere, E y Hinofotis, F (Eds) *Concepts in Language Testing: Some Recent Studies*, Washington (DC): TESOL, 35–49.

Consejo de Europa (1949). *Statutes of the Council of Europe*, documento digital obtenido de https://rm.coe.int/1680306052

Consejo de Europa (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Consejo de Europa (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*, Madrid: Secretaría General Técnica del Ministerio de Educación, Cultura y Deporte y ANAYA.

Consejo de Europa (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Estrasburgo: Consejo de Europa.

Consejo de Europa (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment — Companion Volume with New Descriptors*, Estrasburgo: Consejo de Europa.

Cortázar, J (1963). *Rayuela*, Buenos Aires: Editorial Sudamericana.

Cortázar, J (2016). *Rayuela*, Barcelona: Penguin Random House Grupo Editorial.

Cronbach, L y Meehl, P (1955). Construct validity in psychological tests, *Psychological Bulletin* 52, 281–302.

Crosby, A (1997). *The Measure of Reality: Quantification and Western Society, 1250-1600*. Nueva York (NY): Cambridge University Press.

Cruz, J (2024). Assessment literacy through the design of analytic scales, en Baker, B y Taylor, L (Eds) *Studies in Language Testing (SiLT) 56. Language Assessment Literacy and Competence Volume 2: Case Studies from Around the World,* Cambridge: Cambridge University Press.

Davies, A *et al*. (1999). *Dictionary of Language Testing*, Cambridge: Cambridge University Press.

Davoudi, M y Hashemi, H (2015). Critical review of the models of reading comprehension with a focus on situation models, *International Journal of Linguistics* 7 (5), 172–187.

De Saussure, F (1994). *Curso de lingüística general*, Madrid: Alianza Editorial.

Dean, J (2012). *Rubrics in Language Assessment*, Honolulu (HI): National Foreign Language Resource Center.

Declaración de Bolonia (1999). *Declaración conjunta de los ministros europeos de educación reunidos en Bolonia el 19 de junio de 1999*, documento digital obtenido de https://ehea.info/media.ehea.info/file/Ministerial_conferences/06/0/1999_Bologna_Declaration_Spanish_553060.pdf

Del Jesus, M (2021). *Inteligencia artificial y datos para la sociedad*, Jaén: Universidad de Jaén.

Di Scuillo, A y Boeckx, C (2011). *The Biolinguistics Enterprise. New Perspectives on the Evolution and Nature of the Human Language Faculty*, Oxford: Oxford University Press.

Dick, P (1968). *Do Androids Dream of Electric Sheep?*, Nueva York (NY): Doubleday & Company.

Dick, P (1993). *Do Androids Dream of Electric Sheep?*, Londres: Harper Collins Publishers.

Ding, N *et al*. (2016). Cortical tracking of hierarchical linguistic structures in connected speech, *Nature Neuroscience* 19 (1): 158–164.

Dizon, G y Tang, D (2020). Intelligent personal assistants for autonomous second language learning: an investigation of Alexa, *JALT (Japanese Association of Language Teachers) CALL (Computer Assisted Language Learning) Journal* 16 (2), 107–120.

Eckes, T (2009). Many-facet Rasch Measurement, en Takala, S (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, 1–52.

Edmett, A *et al*. (2023). *Artificial Intelligence and English Language Teaching: Preparing for the future*, Londres: British Council.

Engelhard, G y Wang, J (2021). *Rasch Models for Solving Measurement Problems*, California (CA): SAGE.

Engelhard, G y Wind, S (2018). *Invariant Measurement with Raters and Rating Scales*, Nueva York (NY): Routledge.

Engelhart, M (1965). A comparison of several item discrimination indices, *Journal of Educational Measurement* 2 (1), 69–76.

Escamilla, K *et al*. (2013). *Biliteracy from the start: Literacy squared in action*, Philadelphia (PA): Caslon Publishing.

Ffrench, A (2003). The development of a set of assessment criteria for Speaking Tests, *Cambridge ESOL Research Notes* 13, 8–16.

Field, A (2014). *Discovering Statistics Using IBM SPSS Statistics*, Londres: SAGE.

Fisher, G (1862). *On the numerical mode of estimating and recording educational qualifications as pursued in the Greenwich Hospital Schools by the Rev. George Fisher, M.A., F.R.S., Principal, in a communication to Edwin Chadwick, Esq., President of the Statistical Section of the British Association*, Greenwich: Henry S. Richardson.

Fulcher, G (2003). *Testing Second Language Speaking*, Nueva York (NY): Routledge.

Fulcher, G y Davidson, F (2012). *The Routledge Handbook of Language Testing*. Nueva York (NY): Routledge.

Galante, A (2018). *Plurilingual or monolingual? A mixed methods study investigating plurilingual instruction in an EAP program at a Canadian* university, tésis doctoral, Universidad de Toronto.

Gierl, M *et al*. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: a comprehensive review, *Review of Educational Research* 87 (6), 1082–1116.

Ginther, A (2013). Assessment of Speaking, en Chapelle, C (Ed) *The Encyclopedia of Applied Linguistics*, Blackwell Publishing, 1–7.

Gómez-Benito, J *et al*. (2010). El sesgo de los instrumentos de medición. Tests justos, *Papeles del psicólogo* 31, 75–84.

González-Davies, M (2020). Developing mediation competence through translation, en Laviosa, S y González-Davies M (Eds) *The Routledge handbook of translation and education*, Londres y Nueva York (NY): Routledge, 434–450.

Goodman, K. (1967). Reading: A psycholinguistic guessing game, *Journal of the Reading Specialist* 6 (4), 126–135.

Green, R (2013). *Statistical Analyses for Language Testers*, Eastbourne: Palgrave Macmillan.

Green, R (2017). *Designing Listening Tests: A Practical Approach*, Culemborg: Palgrave Macmillan.

Griffiths, R (1992). Speech rate and listening comprehension: further evidence of the relationship, *TESOL (Teaching of English to Speakers of Other Languages) Quarterly* 26 (2), 385–390.

Hambleton, R *et al*. (1991). *Fundamentals of Item Response Theory*, Newbury Park (CA): SAGE.

Hambleton, R y Jones, R (1993). Comparison of Classical Test Theory and Item Response Theory and their applications to test development, *Educational Measurement: Issues and Practice* 12 (3), 38–47.

Harari, Y (2015). *Sapiens: a Brief History of Humankind*, Nueva York (NY): Harper.

Hasrol, S *et al*. (2022). A systematic review of authenticity in second language assessment, *Research Methods in Applied Linguistics* 1 (3), 1–13.

Heaton, J (1979). *Writing English Language Tests*, Londres: Longman.

Heródoto (1992). *Historia. Libro Segundo. Euterpe*, Madrid: Biblioteca Clásica Gredos.

Hills, P y Argyle, M (2002). The Oxford Happiness Questionnaire: a compact scale for the measurement of psychological well-being, *Personality and Individual Differences* 33 (7), 1073–1082.

IBM Corporation (2023). *IBM SPSS Statistics for Windows* (versión 29.0.2.0), programa informático, Armonk (NY): IBM Corporation.

Instituto Cervantes (2014a). *Guía del examen DELE B1*, documento digital obtenido de https://examenes.cervantes.es/es/dele/preparar-prueba

Instituto Cervantes (2014b). *Guía del examen DELE B2*, documento digital obtenido de https://examenes.cervantes.es/es/dele/preparar-prueba

Instituto Cervantes (2019). *Guía del examen DELE A1*, documento digital obtenido de https://examenes.cervantes.es/es/dele/preparar-prueba

Instituto Cervantes (2022). *Guía del examen DELE A2*, documento digital obtenido de https://examenes.cervantes.es/es/dele/preparar-prueba

International Kendo Federation (2017). *The Regulations of Kendo Shiai and Shinpan*, documento digital obtenido de https://www.ekf-eu.com/documents/202107-Regulations-of-Kendo-Shiai-and-Shinpan.pdf.

Jakobson, R (1962). *Selected Writings*, La Haya: Mouton & Co.

Jentges, S *et al*. (2023). Dutch for young speakers of German – A workshop on receptive multilingualism through cultural and linguistic landscaping, en North, B *et al*. (Eds) *Enriching 21st century language education: The CEFR companion volume, examples from practice*, Estrasburgo: Consejo de Europa, 95–108.

Jurado-Núñez, A *et al*. (2013). Distractores en preguntas de opción múltiple para estudiantes de medicina: ¿cuál es su comportamiento en un examen sumativo de altas consecuencias?, *Investigación en Educación Médica* 2 (8), 202–210.

Kaftandjieva, F (2004). Standard Setting, en Takala, S (Ed) *Reference Supplement to the* Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment *(Section B)*, Estrasburgo: Consejo de Europa, 1–43.

Kane, M (1992). An argument-based approach to validity, *Psychological Bulletin*, 112 (3), 527–535.

Kane, M (1994). Validating the performance standards associated with passing scores, *Review of Educational Research* 64 (3), 425-461.

Kane, M (2001). Current concerns in validity theory, *Journal of Educational Measurement* 38, 319–342.

Kane, M (2006). Validation, en R Brennan (Ed), *Educational measurement*. Westport (CT): American Council on Education/Praeger.

Kazu, I y Kuvvetli, M (2023). The influence of pronunciation education via artificial intelligence technology on vocabulary acquisition in learning English, *International Journal of Psychology and Educational Studies* 10 (2), 480–493.

Keenan, J *et al*. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension, *Scientific Studies of Reading* 12, 281–300.

Keenan, J *et al*. (2009). Assessment and etiology of individual differences in reading comprehension, en Wagner, R *et al*. (Eds) *Beyond Decoding. The Behavioral and Biological Foundations of Reading Comprehension*, Nueva York (NY): The Guilford Press.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items, *Journal of Educational Psychology* 30 (1), 17–24.

Kelley, T *et al*. (2002). Item discrimination indices, *Rasch Measurement Transactions* 16 (3), 883–884.

Kerlinger, F y Lee, H (2000). *Foundations of Behavioral Research*, Fort Worth (TX): Harcourt College Publishers.

Kitaura, F (2020). *Aikido: Method for Flowing Movements*, Amazon Science and Arts: Torrazza Piemonte.

Knoch, U (2009). *Diagnostic Writing Assessment*, Frankfurt: Peter Lang.

Knoch, U (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from?, *Assessing Writing* 16: 81–96.

Knoch, U y Sitajalabhorn W (2013). A closer look at integrated writing tasks: Towards a more focused definition for assessment purposes, *Assessing Writing* 18 (4), 300–308.

Korte, M (2020). The impact of the digital revolution on human brain and behavior: where do we stand?, *Dialogues in Clinical Neuroscience* 22 (2), 101–111.

Lado, R (1961). *Language Testing*, Bristol: Longman.

Lenneberg, E (1967). *Biological Foundations of Language*, Nueva York (NY): Wiley.

Lenneberg, E (1975). *Fundamentos biológicos del lenguaje*, Madrid: Alianza Editorial.

Lewkowicz, J (1996). Authentic for whom? Does authenticity really matter?, en Huhta, A *et al*. (Eds) *Current developments and alternatives in language assessment*, 165–185.

Lewkowicz J (2000). Authenticity in language testing: some outstanding questions, *Language Testing* 17 (1), 43–64.

Linacre, J (1989). *Many-facet Rasch Measurement*, Chicago (IL): MESA.

Linacre, J (1994). Sample size and item Calibration stability, *Rasch Measurement Transactions* 7 (4), 328.

Linacre, J (1997). *Guidelines for rating scales and Andrich Thresholds. MESA Research Note #2*, documento digital obtenido de https://www.rasch.org/rn2.htm

Linacre, J (1999a). Category disordering (disordered categories) vs. threshold disordering (disordered thresholds), *Rasch Measurement Transactions* 13 (1), 675.

Linacre, J (1999b). Investigating rating scale category utility, *Journal of Outcome Measurement* 3 (2), 103–122.

Linacre, J (2002). Optimizing rating scale category effectiveness, *Journal of Applied Measurement* 3 (1), 85–106.

Linacre, J (2024a). *Facets Rasch Measurement* (versión 4.1.6), programa informático, Chicago (IL): Winsteps.com.

Linacre, J (2024b). *Winsteps Measurement* (versión 5.7.2), programa informático, Chicago (IL): Winsteps.com.

Linden, W. van der y Hambleton, R (1997). *Handbook of Modern Item Response Theory*, Nueva York (NY): Springer.

Liu, S y Hung, P (2016). Teaching pronunciation with computer assisted pronunciation instruction in a technological university, *Universal Journal of Educational Research* 4 (9), 1939–1943.

Lloyd-Jones, R (1977). Primary trait scoring, en Cooper, C y Odell, L (Eds) *Evaluating Writing*, Nueva York (NY): National Council of Teachers of English, 33–39.

Loevinger, J (1957). Objective tests as instruments of psychological theory, *Psychological Reports* 3, 635–694.

Lord, F (1980). *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale (NJ): LEA.

MacNeilage, P (2008). *The Origin of Speech*, Nueva York (NY): Oxford University Press.

Malmberg, B (1974). *La América hispanohablante. Unidad y diferenciación del castellano*, Madrid: Ediciones Istmo.

Martínez, A (2011). La evaluación de lenguas. Garantías y limitaciones, Granada: Octaedro Andalucía.

Martin, A y Doumas, L (2017). A mechanism for the cortical computation of hierarchical linguistic structure, *PLOS (Public Library of Science) Biology* 15 (3), 1–23.

Martins, P y Boeckx, C (2016). What we talk about when we talk about biolinguistics, *Linguistic Vanguard 2016*, 1–15.

Massimini, M y Tononi, G (2018). *Sizing up consciousness*, Oxford: Oxford University Press.

Masters, G (1982). A Rasch model for partial credit scoring, *Psychometrika*, 47, 149–174.

McNamara, T (1996). *Measuring Second Language Performance*, Nueva York (NY): Longman.

McNamara, T y Knoch, U (2012). The Rasch wars: the emergence of Rasch measurement in language testing, *Language Testing* 29 (4), 555–576.

MCU (*Magna Charta Universitatum*) (1988). *Magna Charta Universitatum*, documento digital obtenido de https://www.magna-charta.org/magna-charta/en/magna-charta-universitatum/mcu-1988

Meneses, J *et al*. (2013). *Psicometría*, Barcelona: Editorial UOC.

Messick, S (1989). Validity, en Linn, R (Ed) *Educational Measurement*, Nueva York (NY): Macmillan, 13–103.

Messick, S (1995). Standards of validity and the validity of standards in performance assessment, *Educational Measurement: Issues and Practice* 14, 5–8.

Mounin, G (1995). *Historia de la lingüística*, Madrid: Editorial Gredos.

Mustafa, F y Robillos, R (2020). Proper sample sizes for English language testing: a simple statistical analysis, *Humanities & Social Sciences Reviews* 8 (4), 442–452.

Muñiz, J (1998). *Teoría Clásica de los Tests*, Madrid: Editorial Pirámide.

Muñiz, J (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems, *Papeles del Psicólogo* 31 (1), 57–66.

NATO Standardization Office (2016). *NATO Standard Language Proficiency Levels*, documento digital obtenido de https://www.natobilc.org/files/ATrainP-5%20EDA%20V2%20E.pdf

Navarrete, M (2022). La audiodescripción como actividad mediadora en el aula de lenguas, en Sánchez, A (Ed) *Mediación en el aprendizaje de lenguas. Estrategias y recursos*, Madrid: Anaya, 41–66.

Newton, P y Shaw, S (2014). *Validity in Educational & Psychological Assessment*, Londres: SAGE.

North, B (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement, *System* 23 (4), 445–465.

North, B (2000). *The Development of a Common Framework Scale of Language Proficiency*, Nueva York (NY): Peter Lang.

North, B (2007). The CEFR illustrative descriptor scales, *Modern Language Journal*, 91 (4), 656–659.

North, B (2016). Developing illustrative descriptors of aspects of mediation for the Common European Framework of Reference (CEFR): a Council of Europe Project, *Language Teaching* 49 (3): 455–459.

North, B y Piccardo, E (2022). The conceptualisation of mediation in the new CEFR, en Katelhön, P y Marečková, P (Eds) *Sprachmittlung im schulischen und universitären Kontext*, Berlin: Frank and Timme, 23–43.

North, B y Piccardo, E (2023). Plurilingualism and assessment: some issues and options, en Melo-Pfeifer, S y Ollivier, C (Eds) *Assessment of Plurilingual Competence and Plurilingual Learners in Educational Settings*, Londres: Routledge, 178–193.

North, B y Schneider, G (1998). Scaling descriptors for language proficiency scales, *Language Testing* 15 (2), 217–262.

Ortega y Gasset, J (1956). *La rebelión de las masa*s, Madrid: Espasa-Calpe.

Orwell, G (1950). *1984*, Nueva York (NY): Signet Books.

O'Sullivan, B (2014). Validity, validation and development: building and operationalizing a comprehensive model, *JLTA Keynote Speech*, 23–33.

Petrill, S (2009). Genes, environments, and the development of early reading skills, en Wagner, R *et al.* (Eds) *Beyond Decoding. The Behavioral and Biological*

Foundations of Reading Comprehension, Nueva York (NY): The Guilford Press, 246–262.

Piccardo, E *et al*. (2022). *Activating linguistic and cultural diversity in the language classroom*, Nueva York (NY): Springer International Publishing.

Piccardo, E y North, B (2019). *The action-oriented approach: A dynamic vision of language education*, Bristol: Multilingual Matters.

Plakans, L (2008). Comparing composing processes in writing-only and reading-to-write test tasks, *Assessing Writing* 13, 111–29.

Plakans, L (2012). Writing integrated items, en Fulcher, G y Davidson, F (Eds) *The Routledge Handbook of Language Testing*, Nueva York (NY): Routledge, 249–261.

Platón (2004). *Crátilo*, Santa Fe: El Cid Editor.

Poulos, G (2022). When did humans first start to speak? How language evolved in Africa, *The Conversation*, documento digital obtenido de https://theconversation.com/when-did-humans-first-start-to-speak-how-language-evolved-in-africa-194372

Prieto, G (2011). Evaluación de la ejecución mediante el modelo Many-Facet Rasch Measurement, *Psicothema* 23 (2), 233–238.

Prieto, G y Delgado, A (2010). Fiabilidad y validez, *Papeles del Psicólogo* 31(1), 67-74.

Purpura, J (2021). A Rationale for using a scenario-based assessment to measure competency-based, situated second and foreign language proficiency, en Masperi, M *et al*. (Eds) *Évaluation des acquisitions langagières : Du formatif au certificatif. MediAzioni* 32, A54–A96.

R Development Core Team (2024). *R* (versión 4.4.0), programa informático, Auckland: GPL.

Rasch, G (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*, Chicago (IL): The University of Chicago Press.

RAE (Real Academia Española) (2024). *Diccionario de la lengua española* (versión 23.7), diccionario digital consultado en https://www.rae.es

Rehbein, J *et al*. (2012). Lingua receptiva (LaRa) – Remarks on the quintessence of receptive multilingualism, *International Journal of Bilingualism* 16 (3), 248–264.

Reise, S *et al*. (2005). Item Response Theory. Fundamentals, applications, and promise in psychological research, *Current Directions in Psychological Science* 14 (2) 95–101.

Richards, J y R Schmidt (2002). *Longman Dictionary of Language Teaching and Applied Linguistics*, Londres: Longman.

Robins, R (1967). *A Short History of Linguistics*, Thetford: Routledge.

Rumelhart, E (2013). Toward an interactive model of reading, en Alvermann, D *et al*. (Eds) *Theoretical Models and Processes of Reading*, Newark (DE): International Reading Association, 719–747.

San Isidoro de Sevilla (1951). *Etimologías*, Madrid: La Editorial Católica.

Scott, R (1982). *Blade Runner*, película, Warner Bros. Pictures.

Servicio Internacional de Evaluación de la Lengua Española (2021). *Guía para la realización del examen SIELE*, documento digital obtenido de https://siele.org/examen

Shivakumar, A *et al*. (2019). AI-enabled language speaking coaching for dual language learners, *IADIS (International Association for Development of the Information Society) International Journal on WWW/Internet* 17 (1), 66–78.

Simmons, D (2004). *Hyperion Omnibus*. Londres: Gollancz

Skinner, B (1957). *Verbal Behavior*. Nueva York (NY): Appleton-Century-Crofts.

Spearman, C (2010). The proof and measurement of association between two things, *International Journal of Epidemiology* 39 (5), 1137–1150.

Spolsky, B (1977). Language testing: Art or science, en Nickel, G (Ed) *Proceedings of the Fourth International Congress of Applied Linguistics. Volume III*, Stuttgart: Hochschulverlag, 7–28.

Spolsky, B (1995). *Measured Words*, Oxford: Oxford University Press.

Spolsky B (2017). History of Language Testing, en Shohamy E *et al*. (Eds) *Language Testing and Assessment. Encyclopedia of Language and Education*, Cham: Springer, 1–11.

Stathopoulou, M (2018). Assessing cross-Language mediation in the Greek National Multilingual Exam suite: the role of test takers' language competence, en Karavas, E y Mitsikopoulou, B (Eds) *Developments in Glocal Language Testing: The Case of the Greek National Language Proficiency Exams*, Londres: Peter Lang, 243–269.

Steinhuber, B (2022). Implementing plurilingual oral exams and plurilingual lessons in Austrian upper secondary vocational colleges, en North, B *et al*. (Eds) *Enriching 21st century language education: The CEFR companion volume, examples from practice*, Estrasburgo: Consejo de Europa, 109–116.

Stenner, A (1994). Specific objectivity – local and general, *Rasch Measurement Transactions* 8 (3), 374.

Štěpánek, L *et al*. (2023). Item difficulty prediction using item text features: comparison of predictive performances across machine-learning algorithms, *Mathematics* 11 (4104), 1–30.

Stevens, S (1946). On the theory of scales of measurements, *Nature* 103 (2684), 677–680.

Swain, M *et al*. (2009). The speaking section of the TOEFL iBT™ (SSTiBT): test-takers' reported strategic behaviors, *TOEFL iBT™ Research Report* 10, 1–116.

Tang, J *et al*. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings, *Nature Neuroscience* 26, 858–866.

Tarrant M, *et al*. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis, *BMC (BioMed Central) Medical Education* 9 (40), 1–8.

Tena, A *et al*. (2021). Detection of bulbar involvement in patients with amyotrophic lateral sclerosis by machine learning voice analysis: diagnostic decision support development study, *JMIR (Journal of Medical Interest Research) Medical Informatics* 9 (3), e21331.

The Jamovi Project (2024). *Jamovi* (versión 2.3.28), programa informático, obtenido de https://www.jamovi.org

The MathWorks Incorporated (2022). *MATLAB* (versión 9.13.0), programa informático, Natick (MA): The MathWorks Incorporated.

Thomsen, G (1945). *Historia de la lingüística*, Barcelona: Editorial Labor.

Thurstone, L (1928). Attitudes can be measured, *American Journal of Sociology* 33, 529–554.

Tono, Y (2019). Coming full circle—From CEFR to CEFR-J and back, *CEFR Journal Research and Practice* 1, 5–17.

Traub, R (1997). Classical Test Theory in Historical Perspective, *Educational Measurement: Issues and Practice 16* (4), 8–14.

Vallejo, I (2020). *El infinito en un junco. La invención de los libros en el mundo antiguo*, Anzos: Siruela.

Vargas, M (2000). *La fiesta del Chivo*, Madrid: Alfaguara.

Verhelst, N (2004). Item Response Theory, en Takala, S (Ed) *Reference Supplement to the* Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment *(Section G)*, Estrasburgo: Consejo de Europa, 1–42.

Wagner, R *et al*. (Eds) (2009). *Beyond Decoding. The Behavioral and Biological Foundations of Reading Comprehension*, Nueva York (NY): The Guilford Press.

Wang, W (1989). Language in China: a chapter in the history of linguistics, *Journal of Chinese Linguistics* 17 (2), 183–222.

Weigle, S (2002). *Assessing Writing*, Cambridge: Cambridge University Press.

Weir, C (2005). *Language Testing and Validation*, Nueva York (NY): Palgrave Macmillan.

Wolfe, E (2004). Equating and item banking with the Rasch model, en Smith, E y Smith, R (Eds) *Introduction to Rasch Measurement*, Maple Grove (MN): JAM Press, 366–390.

Wolfe, E y Smith, E (2007a). Instrument development tools and activities for measure validation using Rasch models: part 1 – Instrument development tools, *Journal of Applied Measurement* 8 (1), 97–123.

Wolfe, E y Smith, E (2007b). Instrument development tools and activities for measure validation using Rasch models: part 2 – Validation Activities, *Journal of Applied Measurement* 8 (2), 204–234.

Wolfram Research Incorporated (2024). *Mathematica* (versión 14.0), programa informático, Champaign (IL): Wolfram Research Incorporated.

Wright, B (1977). Solving measurement problems with the Rasch Model, *Journal of Educational Measurement* 14 (2), 97–116.

Wright, B (1993). Logits?, *Rasch Measurement Transactions* 7 (2), 288.

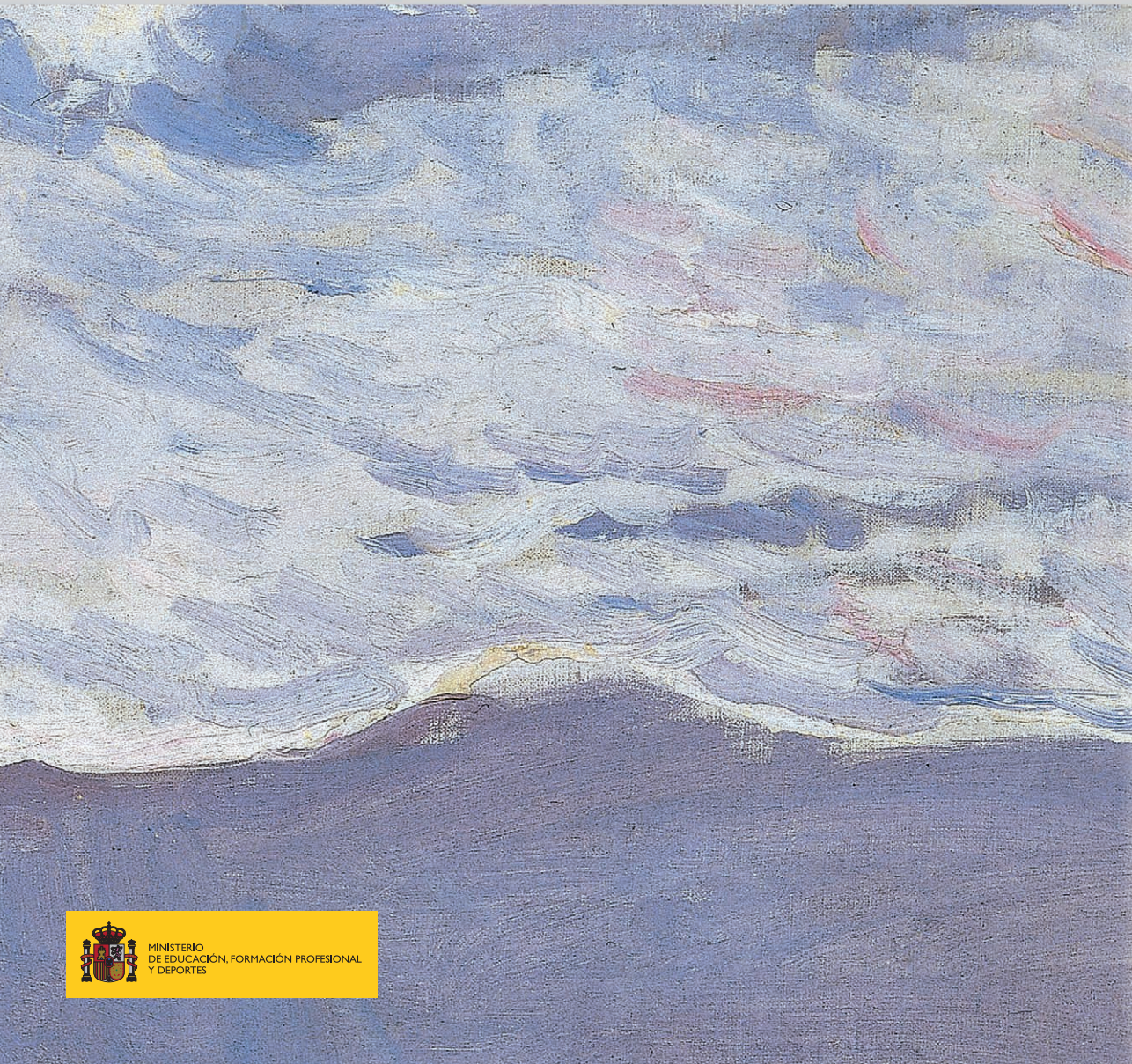
Wright, B y Tennant, A (1994). Sample size again, *Rasch Measurement Transactions* 9 (4), 468.

Wright, B y Stone, M (1978). *Best Test Design. Rasch Measurement*, Chicago (IL): MESA Press.

Wright, B y Masters, G (1982). *Rating Scale Analysis. Rasch Measurement,* Chicago (IL): MESA Press.

Este libro acerca al lector hispanohablante las experiencias que han transformado la evaluación de lenguas durante los últimos treinta años [...]. Como el autor explica en su introducción, esta obra es, en realidad, dos libros: un texto que es un placer leer, que aporta muchas ideas, y un manual de referencia al que acudir mientras se prepara un examen de dominio de lengua.

Brian North



MINISTERIO
DE EDUCACIÓN, FORMACIÓN PROFESIONAL
Y DEPORTES