DISEÑO DE PRUEBAS OBJETIVAS

Instrumentos de evaluación educativa #01

Maite Burruezo Ordóñez,
Javier Cortés De las Heras,
Víctor Martínez Soriano y
Ana Paula Moreno Agud

Diseño de pruebas objetivas. Cuadernos de instrumentos de evaluación #01

Autores: Burruezo Ordóñez, M.T., Cortés De las Heras, J., Martínez Soriano, V.M., Moreno Agut, A.P.

Diseño y maquetación: Perro-Ballena Productions

Imagen de portada: Afonso Lima

2013

2ª revisión 2014

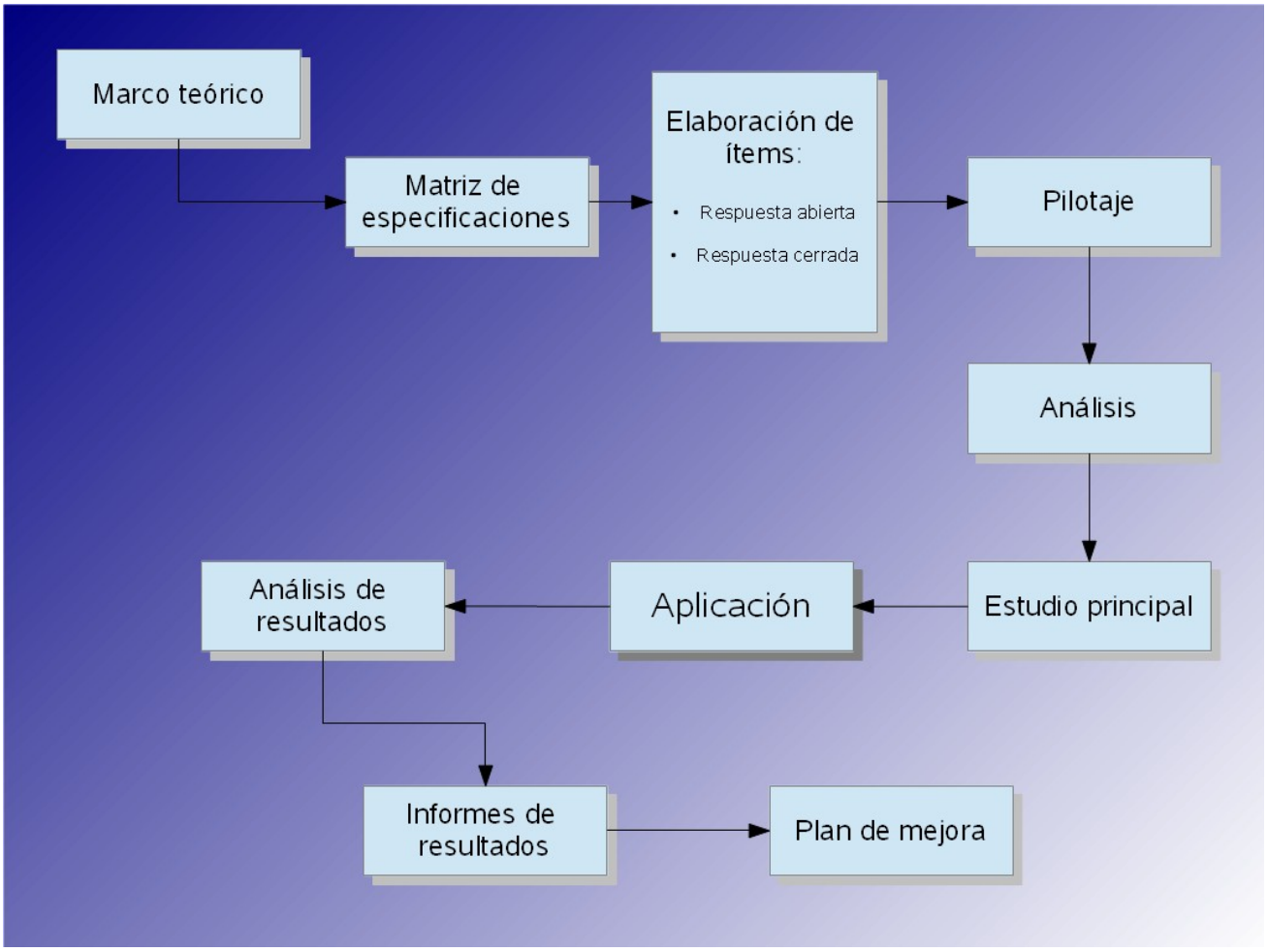
Parte del material de este manual ha sido cedido por el Instituto Nacional de Evaluación Educativa. Ministerio de Educación, Cultura y Deporte.

Índice de contenidos

0. Introducción

Este documento se centra en la exposición de algunos métodos y técnicas para el desarrollo de pruebas objetivas, aunque de forma más precisa prefiramos referirnos a ellas como **cuestionarios de rendimiento**. Esperemos que sean de utilidad para los profesionales de la educación que deseen tener o participar del apasionante proceso de realización de una prueba de carácter externo y también aplicarlo a la elaboración de **pruebas de clase**. De hecho, sería un buen ejercicio comenzar a dotar de calidad técnica las pruebas que solemos realizar a nuestros alumnos, especialmente aquellas que requieren una respuesta escrita.

En el esquema siguiente, se representan las **fases y etapas** de la construcción de un instrumento de este estilo (una de tantas formas de representarlas) desde el diseño de la matriz de especificaciones hasta la consecución de resultados analizables y los consecuentes planes de mejora que deberían implementarse.



1. Marco teórico

Finalidad

Casi todas las pruebas externas, las que realiza la OCDE en el marco de PISA o la IEA: TIMSS o PIRLS, por citar las más conocidas, conllevan instrumentos de medida de la competencia que son cuestionarios de rendimiento. Éstos instrumentos no pueden desarrollarse adecuadamente sin un marco teórico. En las pruebas escritas que planifiquemos en clase quizá no necesitemos un marco teórico al uso para diseñar cuestionarios de rendimiento, pero es interesante conocer qué es y qué finalidad tiene.

El **marco teórico** de una evaluación recoge toda la información necesaria para llevar a cabo desde el diseño de la prueba y del instrumento, hasta la administración del instrumento y ejecución de la prueba por parte del alumnado. Y de ello depende el análisis final que se pueda hacer del rendimiento mostrado por el alumnado. Haciendo un símil, podríamos decir que es como el proyecto en el que se basa el arquitecto para construir el edificio, en él se incluye todo, desde la cimentación al tejado, y no solo su estructura sino que también se da información de instalaciones y otros muchos detalles. En definitiva, no puede existir un buen diseño de instrumentos sin un marco teórico definido previamente.

Posiblemente, la parte más visible de un cuestionario de rendimiento son las preguntas (ítems) que lo configuran y son éstas, en muchas ocasiones, causa de crítica por ser consideradas excesivamente fáciles o difíciles. Esto es debido a que se desconoce lo que pretenden medir, y por ello **es importante no descontextualizar una pregunta del conjunto total de dicho instrumento**.

La **finalidad** de una prueba depende de los **objetivos** a conseguir y, en función de estos, se diseñan los **instrumentos que permiten medir,** con la mayor precisión posible, aquello que queremos analizar. Para ello es necesario establecer:

- los dominios de contenido
- la vinculación con el currículo oficial
- el alumnado objetivo
- la duración de la prueba y las partes de las que consta
- la población final de alumnado evaluada a partir de la que se sacarán los resultados.

Igualmente se hace imprescindible conocer cómo se desean analizar los resultados y las escalas de rendimiento que se emplearán.

Son, por tanto, **pruebas fundamentadas** en algunos casos en **competencias** y en otros en **áreas curriculares**. En el caso de las competencias, como PISA, buscan medir las destrezas y habilidades que los alumnos de esa edad tienen para enfrentarse a determinadas situaciones. En definitiva, se pretende analizar la capacidad de respuesta y reacción frente a situaciones planteadas en las que el alumno o alumna, además de los conocimientos adquiridos en el aula, también debe poner en práctica otros que las diversas situaciones en las que se ha desarrollado su vida le han ido aportando (aprendizajes en la familia, el grupo de amigos, realización de actividades deportivas, vacaciones, etc.).

A la hora de **planificar una prueba escrita** de estas características hemos de tener claro:

- la edad, o el nivel al que se le va a realizar,
- el objetivo (diagnosticar, evaluar, etc),
- establecer el contenido de la prueba,
- las preguntas o ítems y el tipo y número que incluirá el cuestionario,

- el límite del tiempo de la prueba,
- las competencias que se van a implementar y cómo se ponderarán.

Matriz de especificaciones

Éstas y otras características son las que se definen en el marco teórico pero sin duda la **herramienta base** que facilita y guía la construcción y la interpretación de pruebas es lo que denominamos **matriz o tabla de especificaciones**. También se denomina tabla curricular de referencia, dado que en ella definimos descriptores según el bloque de contenidos al que apuntan y a la competencia con la que se relacionan. Definamos los **conceptos implicados:**

- **Contenidos**: Declaraciones conceptuales o procedimentales que forman el cuerpo de enseñanza de una disciplina.
- **Procesos**: Niveles de complejidad, cognitiva, en la resolución de una tarea.
- **Descriptores:** Formulaciones sintéticas de los micro-conocimientos que habrán de ser medidos por los ítems que se elaboren. De cada uno de ellos se pueden construir varios ítems.

Para elaborar una matriz necesitamos:

- Establecer los **contenidos.**
- Establecer los **procesos cognitivos** o grados de adquisición de la competencia.
- Determinar los **descriptore**.
- Determinar el **peso relativo** de cada casilla, donde se relacionan los contenidos con los procesos.
- **Elaborar ítems** específicos para cada descriptor.

Pongamos un **ejemplo para poder entender mejor lo que es una matriz de especificaciones**. En ella se clasifican una serie de descriptores que apuntan biunívocamente a un grado de adquisición de la competencia y a la vez al bloque de conocimientos con el que se vincula.

	Niveles cognitivos (grado de adquisición de la competencia)		
Contenidos	**Proceso I**	**Proceso II**	**Proceso III**
Contenido A	Descriptores	Descriptores	Descriptores
Contenido B	Descriptores	Descriptores	Descriptores
Contenido C	Descriptores	Descriptores	Descriptores
Contenido D	Descriptores	Descriptores	Descriptores

En la tabla se observa como el conjunto de **descriptores** por ejemplo los representados **en rojo**, se relacionan con un proceso cognitivo inicial (Proceso I), o lo que podríamos definir como un grado de adquisición de la competencia bajo, ligado a procesos de reproducción de contenidos elementales, en este caso a contenidos de tipo C.

De igual manera, si nos fijamos en el conjunto de descriptores del Proceso II, de color **azul**, estaríamos vinculando éstos con un grado de adquisición de competencia medio,

con los contenidos de tipo B, donde el alumno necesita estrategias que permiten acciones de conexión de diversos conocimientos. Estamos por tanto frente a un grado mayor de competencia.

Si observamos el color **verde** del ejemplo estaríamos frente a descriptores que medirían un grado elevado de adquisición de la competencia (Proceso III) y como en los casos anteriores estos estarían vinculados a un grupo de contenidos concretos, pero las acciones que desarrollaría el alumno estarían basadas en procesos de reflexión, lo cual supone un mayor grado de complejidad en la resolución de las tareas planteadas para dicho proceso.

Matriz de especificaciones: contenidos y procesos

Todos y cada uno de estos bloques en los que hemos subdividido las matemáticas implican **procesos I, II y III,** de menor a mayor dificultad según sea mayor la adquisición de la competencia o proceso cognitivo. Por ejemplo, nombremos éstos como **Reproducción** (grado bajo de adquisición de la competencia), **Conexión** (grado medio de adquisición de la competencia) y **Reflexión** (grado elevado de adquisición de la competencia).

No olvidemos que estamos diseñando una prueba de evaluación, por lo que **una pregunta o ítem estará vinculada a un descriptor**, que como hemos visto implicará un proceso cognitivo determinado y guardará relación con un bloque de contenido concreto. Por tanto, estos **descriptores** generalmente tienen la forma de **criterio de evaluación**, de hecho son eso mismo, dado que con ellos intentamos medir el rendimiento del alumno.

Como **con cada ítem se mide un único descriptor**, en realidad podríamos hablar de microcriterios de evaluación, al tener que dar cada uno de ellos una información única y específica. Se trataría, por tanto, de una especificación de los criterios de evaluación del currículo que, en el caso del currículo oficial de la Comunitat Valenciana para Ed. Primaria, han tomado el nombre de indicadores de logro.

Cuadro de relaciones de Educación Primaria (continuación)

Bloques contenidos		Reproducción		Conexión		Reflexión	
		Acceso e identificación	Compresión	Aplicación	Análisis y valoración	Síntesis y creación	Juicio y regulación
Geometría	Criterios de evaluación: 5, 6 y 8	• Obtener información puntual de una representación espacial. • Reconocer formas y cuerpos geométricos. • *Lng.:* Localizar y recuperar información explícita.	• Comprender situaciones geométricas de la vida cotidiana. • Conocer las propiedades básicas de cuerpos y figuras planas. • *Lng.:* Captar el sentido global y algunas informaciones específicas.	• Utilizar las nociones básicas de los movimientos geométricos. • Describir situaciones geométricas de la vida cotidiana. • Utilizar los movimientos en el plano para emitir y recibir informaciones sobre situaciones cotidianas. • Reproducir manifestaciones artísticas que incluyan simetrías y traslaciones. • Identificar manifestaciones artísticas que incluyan simetrías y traslaciones. • Describir formas y cuerpos geométricos del espacio (polígonos, círculos, cubos, prismas, cilindros y esferas). • Clasificar tanto figuras como cuerpos. • Aplicar los conocimientos adquiridos. • *EF.:* Girar sobre el eje longitudinal y transversal diversificando las posiciones segmentarias (utilizarlos en las actividades cotidianas).	• Clasificar tanto figuras como cuerpos con criterios libremente elegidos. • Resolver problemas relacionados con el entorno que exijan cierta planificación, aplicando los contenidos básicos de geometría. • *E. Art.:* Interpretar el contenido de imágenes y representaciones del espacio.	• Describir, en situaciones de la vida cotidiana, una representación espacial (croquis de un itinerario, plano de una pista...) tomando como referencia objetos familiares. • Tener capacidad de orientación y representación espacial, teniendo en cuenta tanto el lenguaje utilizado como la representación en el plano de objetos y contextos cercanos. • Utilizar estrategias personales para la resolución de problemas. • Utilizar más de un procedimiento en la resolución. • Expresar de forma escrita y ordenada el proceso. • *C. Medio:* Comunicar de forma escrita los resultados acompañados de tablas, gráficos, etc. • *Lng.:* Interpretar e integrar las ideas propias con la información contenida en los textos adecuados al nivel de edad. • *Lng.:* Redactar, reescribir y resumir textos en situaciones cotidianas y escolares de forma ordenada y adecuada, utilizando la planificación.	• Valorar las diversas expresiones artísticas. • Valorar la utilización de propiedades geométricas (alineamiento, paralelismo, perpendicularidad..) como elementos de referencia para describir situaciones espaciales.

Ilustración 1: Tabla de especificaciones de Competencia Matemática-Geometría, 4º Ed. Primaria.

Fuente: Evaluación General de Diagnóstico 2009. Marco de la evaluación. Ministerio de Educación.

Matriz de especificaciones: ítems, pilotaje

Una vez que se han identificado todos los criterios que configuran la matriz, es necesario **generar preguntas o ítems** que estén vinculados con ellos. Además es necesario hacer un elevado número de preguntas de forma que se puedan **pilotar** (ponerlas en práctica para comprobar que lo que se ha establecido en la matriz de especificaciones es contrastado por la realidad de lo que ocurre al enfrentar al alumnado a las mismas). Es decir, **si una pregunta ha sido creada para medir un proceso cognitivo, que realmente lo mida**. Seguramente de las preguntas que se han realizado, algunas se eliminarán, normalmente por un mal funcionamiento o anomalías observadas al aplicarlas con un grupo de alumnos, una mala redacción del enunciado, error en algunas de las respuestas, duplicidad de claves, errores tipográficos, que la dificultad esté en la forma de preguntar y no en lo que se pregunta, etc.

El proceso de **pilotaje** se realiza con alumnos que no tendrán dificultad en responder a lo que se les pregunta, de manera que lo que se analiza es la validez y viabilidad del ítem. Por ejemplo, si diseñáramos una prueba para 2º de ESO lo podríamos pilotar con alumnos de 3º de ESO en el momento de inicio del curso escolar (octubre o noviembre).

Al final, el instrumento quedará configurado de forma que exista un conjunto de **preguntas que cubran todos los procesos cognitivos y conocimientos**.

Debemos tener en cuenta que **cuando un profesor hace una prueba escrita utilizando el cuestionario**, normalmente establece un par de preguntas relativamente sencillas para tratar de que el mayor número de alumnos se aproxime al aprobado. De igual modo incluye una o dos preguntas para diferenciar aquellos alumnos que destacan y el resto de preguntas podríamos considerarlas de dificultad intermedia. En estas pruebas pasa algo parecido, por eso descontextualizar una pregunta del conjunto de la prueba no tiene mucho sentido. Hay preguntas muy fáciles, para conocer el alumnado que al menos llega a este nivel mínimo y quien no lo alcanza. De igual modo interesa saber qué porcentaje destaca de entre todo el alumnado, de ahí la existencia de algunas preguntas de cierta complejidad, siendo el grueso de dificultad intermedia que es donde normalmente se sitúa la mayor parte del alumnado.

Tabla 1.11. Pesos de la matriz de especificaciones en la Competencia matemática

Bloques de contenidos	Procesos						Pesos (%)
	Reproducción		Conexión		Reflexión		
	Acceso e identificación	Comprensión	Aplicación	Análisis y Valoración	Síntesis y Creación	Juicio y Regulación	
Números y operaciones							35 ± 5
La medida							20 ± 5
Geometría							25 ± 5
Tratamiento de la información, azar y probabilidad							20 ± 5
Pesos (%)	**10 ± 5**	**15 ± 5**	**25 ± 5**	**20 ± 5**	**20 ± 5**	**10 ± 5**	**100**

Ilustración 2: Pesos de la matriz de especificaciones. Competencia matemática, 4º Ed. Primaria.

Fuente: Evaluación General de Diagnóstico 2009. Educación Primaria. Cuarto curso. Informe de resultados, p. 24. Ministerio de Educación.

En las matrices figura el **porcentaje de ítems** para cada dominio (peso relativo), referido tanto a los contenidos como a los procesos cognitivos.

Obsérvese que los **descriptores están vinculados a dominios de contenido y dominios cognitivos.**

Tabla 1.12. Matriz de especificaciones en la Competencia matemática

Bloques de contenidos	Procesos					
	Reproducción (16%)		Conexión (71%)		Reflexión (13%)	
	Acceso e identificación (7%)	Comprensión (9%)	Aplicación (34%)	Análisis y valoración (37%)	Síntesis y creación (10%)	Juicio y regulación (3%)
Números y operaciones (36%)		M029	M011 M019 M024 M025 M027 M034 M037 M057 M067 M072 M073 M074	M026 M030 M031 M032 M035 M036 M060 M069 M078 M079	M009 M033 (2 puntos) M038 (2 puntos)	M002
La medida: estimación y cálculo de magnitudes (21%)		M050 M051 M052 M053 M054 M055 M056	M003 M041 M064 M066 M068	M004 M006 M012	M065	
Geometría (21%)	M045 M049 M058	M046 M047 M061	M005 M013 M039 M040 M044 M059	M048	M043 M063	
Tratamiento de la información, azar y probabilidad (22%)	M008 M075	M020 M070 M071 M076	M007 M021 M077	M014 M015 M016 M017 M018	M010 M001 M022 (2 puntos)	M023 (2 puntos)

Ilustración 3: Ítems distribuidos en el instrumento final y pesos definitivos. Competencia matemática – 4º Ed. Primaria.

Fuente: Evaluación General de Diagnóstico 2009. Educación Primaria. Cuarto curso. Informe de resultados, p. 25. Ministerio de Educación.

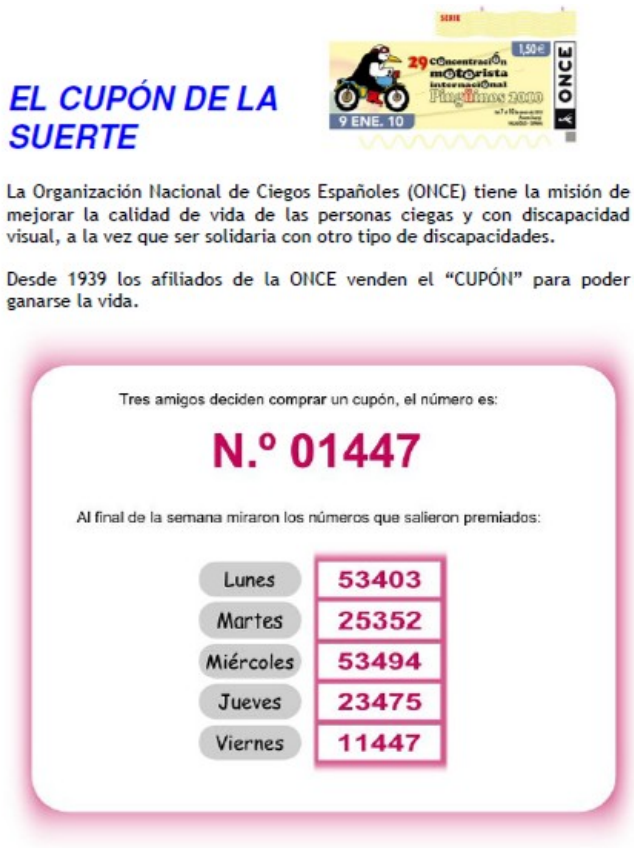
2. Los estímulos

Las preguntas de un cuestionario de rendimiento, por lo general, no se presentan aisladas sino formando grupos bajo una presentación textual y/o gráfica común llamada **estímulo**, que presenta al alumno una situación cotidiana como las que se pueden encontrar en la vida real.

En el caso de la comprensión lectora normalmente son **textos**, pero también pueden ser **trípticos, esquemas, documentos oficiales**, etc. Esta presentación permite minimizar el efecto de los cambios de contexto de las preguntas y también posibilita que el estímulo común pueda ser mejor explotado. Se pueden incluir gráficos, planos, medidas, cómics, etc.

El estímulo o contexto de partida puede crearse a partir de varios elementos y por tanto podemos diseñarlos para que se adapten a las preguntas que vamos a realizar en referencia a los descriptores.

El alumnado actual está inmerso en un mundo de información a través de Internet, por lo que gran cantidad de información a la que se enfrentan está en ese entorno. Podemos emplearlo como un recurso muy rico de estímulos sin olvidar otros como folletos informativos, textos o anuncios de periódicos, fotografías de situaciones reales, etc.



Fuente: Materiales liberados. Evaluación de Diagnóstico 2012. Competencia Matemática 4º Ed. Primaria. Comunitat Valenciana.

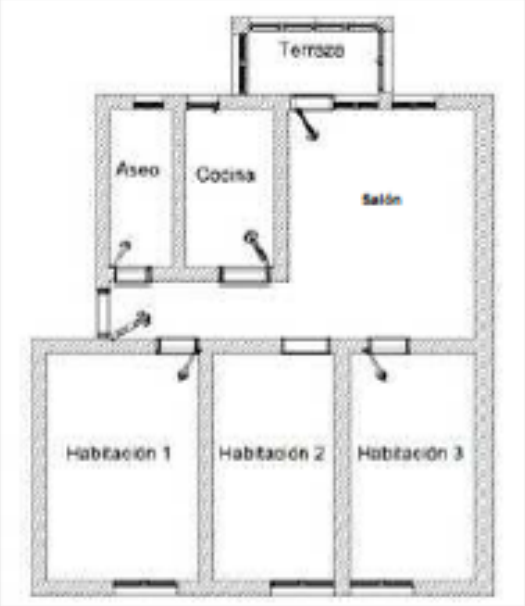
LA VIVIENDA

Este acogedor piso de 90 m^2 construidos (76,5 m^2 útiles) se encuentra en una zona muy tranquila, sin ruidos y con zonas ajardinadas.

El edificio se encuentra a cinco minutos del centro, lo que permite desplazarse sin tener que utilizar el coche. Tampoco es demasiado complicado encontrar un hueco para aparcar y por la zona hay varios garajes donde alquilar o comprar una plaza si se desea.

En cuanto al ambiente que lo rodea, es una comunidad muy tranquila, con buenas relaciones entre vecinos. El precio de la comunidad es bajo, 15 €/mes.

El precio de venta que figura en la inmobiliaria es de 150.000 €.



Fuente: Materiales liberados. Evaluación de Diagnóstico 2012. Competencia Matemática 2º ESO. Comunitat Valenciana.

TREKKING EN ÁFRICA

Trekking en el norte del Drankensberg: Sudáfrica / Lesotho

Ficha informativa

DESCRIPCIÓN GENERAL

- El trekking por el norte del Drakensberg incluye atravesar la escarpadura norte del Drakensberg a gran altitud. La ruta, de unas 40 millas (65 km) aproximadamente, se extiende a lo largo de la frontera entre Sudáfrica y Lesotho, y precisa 5 agotadores días para completarse. El trekking está lleno de momentos espectaculares, como las impresionantes vistas desde más allá del Anfiteatro hasta el Diente del Diablo, a medida que se va en dirección a la Escalera Metálica, y la salida del sol vista desde el Mponjwane, para la que bien merece la pena poner el despertador.
- Punto de partida: aparcamiento El Centinela, Parque Nacional Real Natal.
- Punto de llegada: hotel El Pico de la Catedral.
- Dificultad y altitud: se trata de un camino de alta montaña en una de las zonas más remotas de la cordillera del Drakensberg. La marcha puede resultar bastante ardua y los días largos. Un buen sentido de la orientación es fundamental para realizar la travesía con seguridad.

ÉPOCA ADECUADA Y DIFERENCIAS ESTACIONALES

- Mejores meses para viajar: abril, mayo, junio o septiembre, octubre, noviembre.
- Clima: los veranos en el Drakensberg pueden ser muy cálidos y muy húmedos. Los inviernos son mucho más secos, pero siempre existe el riesgo de precipitaciones, probablemente en forma de nieve en las zonas altas. En la primavera y el otoño las temperaturas diurnas son ideales (entre 60°F/15°C y 70°F/20°C), pero por la noche caen frecuentemente por debajo del punto de congelación.

Temperatura y Precipitaciones												
Temperatura máxima (media diaria)												
(°F)	72	70	70	66	63	60	60	63	66	68	70	70
(°C)	22	21	21	19	17	15	15	17	19	20	21	21
Temperatura mínima (media diaria)												
(°F)	55	55	54	48	46	41	41	43	46	48	52	54
(°C)	13	13	12	9	8	5	5	6	8	9	11	12
Precipitaciones (media mensual)												
(Pulgadas)	9,3	8,5	7,7	3,1	1,1	0,6	0,5	1,3	2,4	4,0	6,5	7,9
(Milímetros)	237	216	196	78	29	14	12	33	62	101	165	201
	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC

Fuente: Materiales liberados. Evaluaciones de Educación Secundaria (INEE), pp. 70-71

Recomendaciones y ejemplos para la selección de estímulos

Utilidad

- Conectar o contextualizar la competencia o área en el mundo real.
- Hace que las preguntas (ítems) tengan un propósito y se alineen con los requisitos del marco teórico.
- Hace que los ítems sean interesantes y más atractivos visualmente para los alumnos que simplemente textos o símbolos.
- El contexto y el realismo hacen la prueba más interesante para los alumnos.
- Permite extraer un número de preguntas a realizar sobre un mismo estímulo o contexto, configurando así una unidad de evaluación.

Criterios para seleccionar un buen estímulo

Debe:

- Ser rico en cuanto a los matices que proporciona la información que contiene, e interesante.
- Presentar un reto óptimo a los estudiantes, ni demasiado difícil ni demasiado fácil.
- Ofrecer la oportunidad de realizar preguntas.
- Ser más o menos accesible y equitativo atendiendo a posibles diferencias en los alumnos.

No debe:

- Plantear retos artificiales o añadidos.
- Ofender o molestar:
  - Situaciones traumáticas (accidentes de coche, violencia),
  - sexo, religión, política u otros asuntos que sean emocionalmente conflictivos,
  - malas conductas, violencia, racismo, inmoralidad o irresponsabilidad,
  - modelos no deseables (inducción al consumo de drogas, alcohol u otras conductas potencialmente peligrosas),
  - lenguaje soez.
- Esperar demasiado o muy poco de los estudiantes.
  - Que se base en conocimiento poco o nada familiar,
  - Para alumnos de más edad que les haga pensar que el estímulo es para niños pequeños.

<u>En textos particularmente:</u>

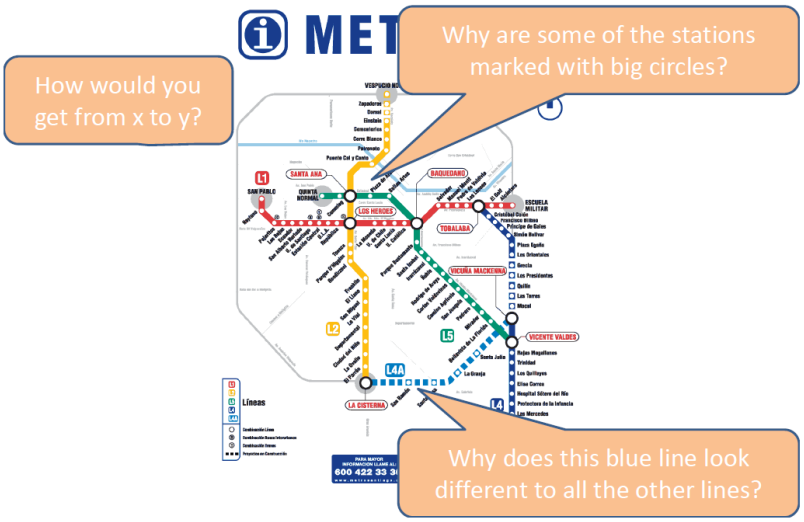
- Textos cuya base sea un conocimiento poco difundido.

- Textos que proporcionan un conocimiento tan general que la mayoría de los alumnos ya posee.
- Textos que son aburridos.

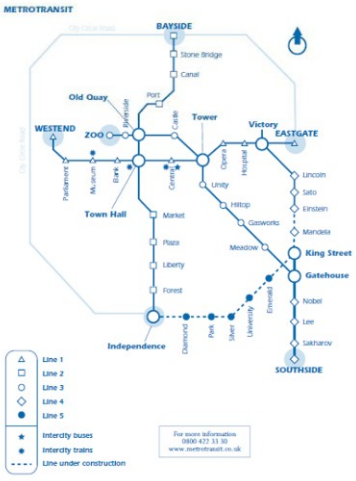
En resumen debemos atender a:

- Grado en el que refleja el mundo real.
- Grado de adecuación cultural.
- Grado de adecuación lingüística.
- Grado de interés para el grupo diana.
- Grado de dificultad.
- Derechos legales (copyright, licencia libre de diversos tipos...).

A non-continuous text (original version)



A non-continuous text (modified for PISA)



Fuente: Evaluaciones PISA.

Dónde se pueden encontrar estímulos

- Noticias, artículos – en papel o en Internet.
- Compras, incluyendo anuncios.
- Objetos reales y su embalaje – comida, bebidas, ingredientes, materiales, jardín, hogar, medicinas, productos de limpieza, etc.
- Cocina y comida – recetas.
- Del mundo de la música o del cine.
- De deportes – rugby, por ejemplo.
- Panfletos y octavillas.
- Mapas y guías.
- Manualidades, aficiones, etc.
- Información financiera.
- Información sobre edificios, jardines, campos de deporte, parques, etc.
- Información sobre vacaciones y viajes.
- Materiales de trabajo, si se ve apropiado.
- Incluso una foto con texto al pie.
- ...

En las pruebas de evaluación de la competencia lingüística, se utilizan como estímulos **textos** (continuos, no continuos o mixtos). En los no continuos, muchos de los ejemplos anteriores podrían caber.

Ejemplo de texto continuo

El xiquet amb el pijama de ratlles

THE BOY IN THE STRIPED PYJAMAS (USA, Gran Bretanya. 2008. 93 min.). **Direcció i guió:** Mark Herman sobre la novel·la de John Boyne. **Intèrprets:** Ansa Butterfield. Jack Scanlon. David Thewlis. Vera Farmiga. Rupert Friend. **Fotografia:** Benoît Delhomme. **Música:** James Horner. **DRAMA.**

A FAVOR

Per Nuria Vidal

El millor que es pot dir d'esta pel·lícula és que és tot el contrari de *La vida és bella*, de Roberto Benigni. El que allí era edulcorant superficialitat per a despertar la llàgrima fàcil, ací es convertix en un profund i terrible horror. No l'horror dels camps d'extermini nazis a què el cine, desgraciadament, ens ha acostumat, sinó l'horror de vore amb uns ulls innocents i nets el que eixos camps eren en realitat. La valentia d'esta història és la de mostrar el contraplà. El cine ens ha ensenyat moltes vegades el que passava en els camps des de la xarxa de filferro cap a dins, però quasi mai el que succeïa fora. Què sentien les famílies dels caps nazis dels camps d'extermini? Com suportaven viure sabent el que hi passava? Pot un xiquet de huit anys entendre què és la Granja i que els grangers vagen sempre vestits amb un pijama? Herman no perd mai el punt de vista de Bruno, el xiquet solitari que busca algú amb qui jugar en eixe món on res és el que sembla. Tot ho vivim a través d'ell. Només al final la pel·lícula deixa Bruno sumit en la foscor, i els altres, perduts per sempre.

EL PITJOR: les odioses comparacions que sorgiran al seu voltant.

EN CONTRA

Per Mirito Torreiro

Encara negant la terrible sentència de Theodor W. Adorno, segons la qual després d'Auschwitz ja no és possible la poesia, no hi ha dubte que quan s'aborda l'exterminació dels jueus en els camps nazis cal fer-ho amb atenció, com fa el jueu Spielberg, amb *La Llista de Schindler*, no una pel·lícula sobre l'Holocaust sinó una història de triomfadors (la definició és de S. Kubrick). D'ací que resulte escandalosa l'opció del, d'altra banda, competent Mark Herman (no opine sobre la novel·la que està en la base del film, que no conec) a l'hora de construir el punt de vista d'este film sobri i ben narrat (i per això molt més sinistre): situant-lo en els ulls del fill d'un comandant nazi, ens fa que patim moltíssim pel destí de la criatura... mentres en els forns crematoris cremen els cadàvers de centenars d'altres víctimes, que no veiem. Només amb açò hi ha prou per a odiar un film tan trampós; però és que hi ha més: una mare que no sap a què es dedica el seu marit, una germaneta disposada a sacrificar-se per Hitler per a, sobtadament, deixar de fer-ho... com si tota Alemanya haguera sigut resistent...

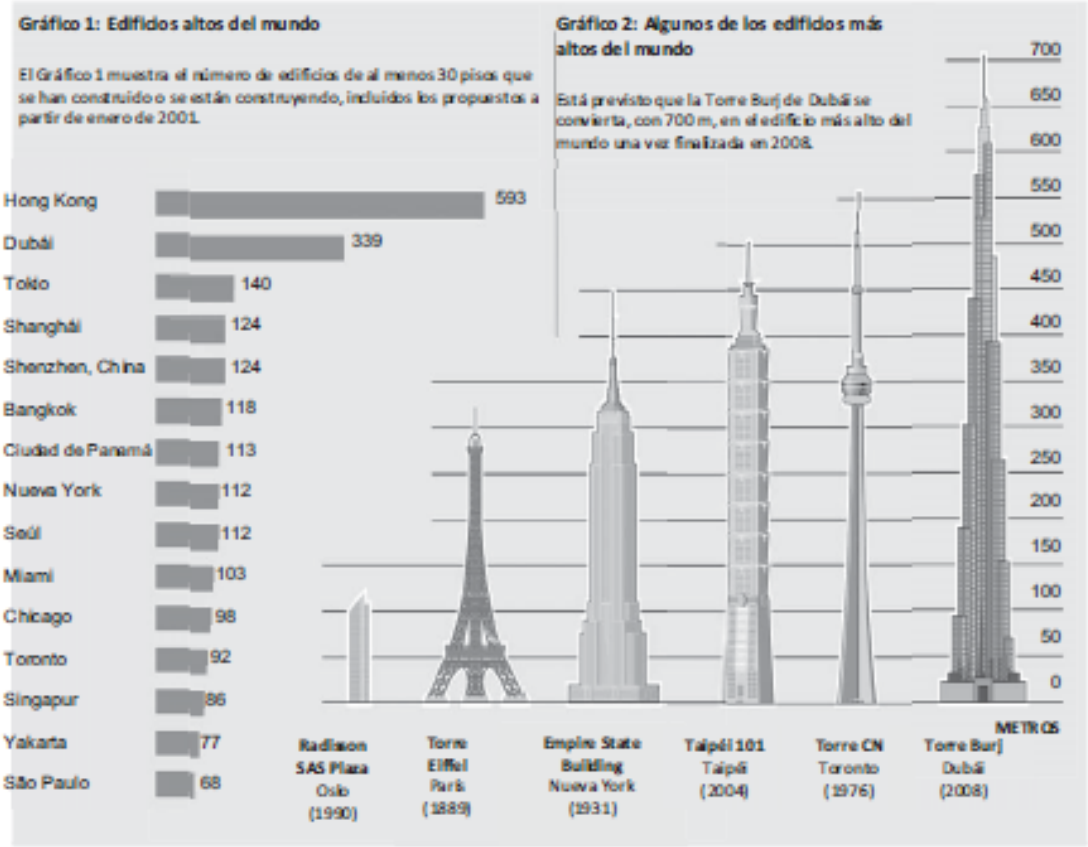
EL MILLOR: la cuidada posada en escena de la pel·lícula.

Fotogramas, Octubre 2008 (Adaptació)

Ejemplo de texto no continuo

EDIFICIOS ALTOS

«Edificios altos» es un artículo de una revista noruega publicado en 2006.



Este texto yuxtapone dos gráficos que guardan cierta relación en cuanto a contenido. Los dos hacen referencia a edificios altos que hay en el mundo: el *Gráfico 1* muestra el número de edificios de este tipo que existen en distintas ciudades, en proyecto o construidos, mientras que en el *Gráfico 2* se presentan algunos de los edificios más altos del mundo. Aunque cada uno de ellos está introducido por un breve fragmento explicativo en prosa, la información fundamental del texto se facilita en los dos gráficos, convirtiendo el formato de texto global en discontinuo. El tipo de texto se corresponde con la descripción, mientras que la situación es educativa, puesto que apareció en una revista para estudiantes. El artículo comienza con una breve introducción que explica el contexto, tanto en lo referente al tiempo (el artículo se publicó en 2006) como al lugar (la revista es noruega). Una de las razones por las que esta

Fuente: Materiales liberados. Evaluaciones de Educación Secundaria (INEE), p. 50.

Aspectos diferenciales en los estímulos de competencia matemática y lingüística

COMPETENCIA MATEMÁTICA

El reto en la competencia matemática

Para realizar buenos estímulos en las pruebas de competencia matemática, se han de proponer aquellos que suponga un reto asequible para los alumnos. Sin ser excesivamente difíciles ni demasiado fáciles. Obviamente para conseguirlo, el estímulo tiene que evaluar los componentes del marco teórico de la competencia matemática – contextos, contenidos y procesos.

Además debe ser:

- Un estímulo auténtico, que generalmente ha de simplificarse.
- Accesible, incluyendo el contexto, el lenguaje y la terminología, pues no es una evaluación de lectura.

COMPETENCIA LINGÜÍSTICA: MAPAS DE TEXTO

Textos de ficción

Las características principales que deben analizarse son:

- Tema
- Hechos principales
- Personajes principales
- Contexto
- Características del lenguaje

Ejemplo: Nunca más comeré tomates

(texto prueba de la Evaluación de Diagnóstico en Competencia en Comunicación Lingüística, 2013)

- *Tema*: alimentación, importancia de comer de todo, ver las cosas desde otro punto de vista.
- *Hechos principales*: Tolola no quiere comer ciertos alimentos; Juan les cambia el nombre original por otro fantástico; Tolola empieza a comer todo lo que no le gustaba; Tolola entre en el juego, y por propia iniciativa crea un nuevo nombre.
- *Personajes principales*: Juan, Tolola (hermanos).
- *Contexto*: En la mesa a la hora de cenar (en un hogar).
- *Características del lenguaje*: vocabulario de alimentos, uso de la metáfora, nombres de lugares (Fuji, Groenlandia, Júpiter, Luna), exclamaciones e interrogantes, sirenas.

Textos de no ficción

Las características principales que deben analizarse son:

- Propósito: exposición, argumentación, crítica...
- Tema: descripción del tema principal del texto.
- Contenido: descripción general del contenido del texto, los rasgos más fundamentales.
- Estructura: mapa de estructura de ese contenido, relaciones jerárquicas en los conceptos...
- Características de la presentación: prosa o no, si tiene ilustraciones y diagramas con anotaciones...
- Características del lenguaje.

COMPETENCIA LINGÜÍSTICA: ANÁLISIS ESTRUCTURAL

Además de los mapas de texto, también ayuda la realización de un análisis estructural de los textos. Ambas técnicas van a facilitar el desarrollo de buenos ítems.

Siguiendo con el ejemplo de la lectura anterior "*Nunca más comeré tomates*", se presenta el análisis de su estructura.

INTRODUCCIÓN

Tolola: enumera los alimentos que no le gustan y no come

Y desde luego nunca jamás comeré tomates

NUDO

Juan: dice que no comerán ninguno de esos alimentos

Metáfora 1: zanahorias → varitas mágicas de **Júpiter**

Metáfora 2: guisantes → gotas verdes de **Groenlandia**

Metáfora 3: puré de patata → trocito de nube del **monte Fuji**

Metáfora 4: varitas de pescado → bocaditos de **mar** – *comida de las sirenas*

Tolola come esos alimentos

DESENLACE

Tolola pide comer tomates

Metáfora 5: tomates → chorros de la **Luna**

¿Qué características conforman una buena unidad de evaluación?

Una unidad de evaluación es el conjunto de estímulo y sus ítems asociados.

Generalmente, aunque varía según la organización que promueve la evaluación y la competencia evaluada, un estímulo tiene asociados un mínimo de 5 ítems. 15 o más ítems, en el caso de los textuales.

Los ítems de un estímulo deberían cubrir todo el rango de dificultad.

Los ítems también deberían cubrir las características del marco teórico.

Cada ítem es independiente de los otros. La independencia es un requisito técnico indispensable.

En estímulos textuales, los ítems han de referirse a las características principales del texto, no a anotaciones o información al margen.

NOTA: En muchas ocasiones, especialmente con un número pequeño de ítems asociado al estímulo, es complicado cubrir todo el rango del marco teórico y dificultad. En este caso, es preferible optar por cubrir el rango de dificultad.

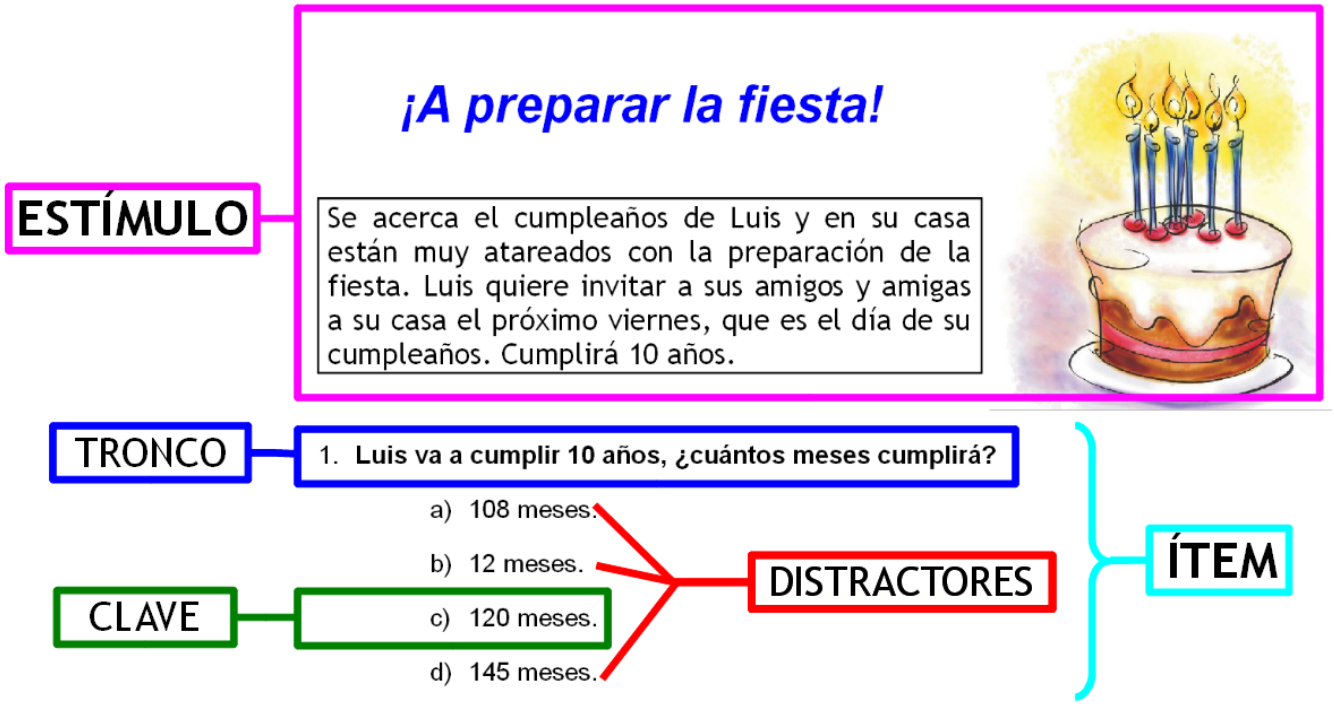
3. Ítems

¿Qué es un ítem?

- Es una unidad de medida.
- Consiste en un reactivo (tronco) y una forma prescrita de respuesta.
- Su objetivo es obtener una conducta observable y claramente interpretable (debe ser capaz de discriminar).
- De la respuesta debe poder inferirse, con un grado suficiente de fiabilidad y validez, un nivel de dominio de la competencia.

Conozcamos cuales son los componentes que configuran un ítem de respuesta cerrada:

- **Tronco:** Parte inicial del ítem en la que se especifica la tarea. Puede ser una pregunta, instrucciones, una frase incompleta, una situación de la vida real, etc.
- **Opciones o alternativas**: Son todas las alternativas de respuesta de un ítem.
- **Clave**, es la respuesta correcta de las alternativas dadas.
- **Distractores**, son las respuestas incorrectas de las alternativas dadas.



Un estímulo trata de ser un elemento de partida que posiciona al alumno en una situación que pudiera ser real. Normalmente estas pruebas vinculan varios ítems o preguntas con el mismo estímulo.

Un objetivo importante en las evaluaciones externas es la comparación a lo largo del tiempo (en las sucesivas ediciones), por lo que sólo se liberan parte de los materiales que configuran la prueba y nunca una prueba completa, de forma que una parte importante

de la prueba se repite de una edición a otra. A estos ítems se les denomina ítems de anclaje.

Tipos de ítems o preguntas

Las preguntas incluidas en un cuestionario de rendimiento pueden ser de diversos tipos. A continuación se muestra una selección de preguntas clasificadas atendiendo a la forma de la respuesta.

Preguntas de respuesta cerrada o estructurada, son las que muestran alternativas de respuestas.

- De respuesta alternativa simple (dicotómicas), combinan diferentes opciones de respuesta binaria (sí / no; verdadero / falso).
- De respuesta de alternativa múltiple. Presentan varias alternativas (A, B, C, D).
- De respuesta de asociación.
- De respuesta de elección múltiple compleja, cuando se combina en un mismo ítem varias frases o alternativas con selección de respuesta binaria (sí / no; verdadero / falso, o asociación).

Pregunta de respuesta cerrada

Respuesta alternativa simple	La bombilla la inventó Thomas Alva Edison. SI/NO
Respuesta alternativa múltiple	¿Quién inventó la bombilla? A. Johannes Gutenberg B. Thomas Alva Edison C. Albert Einstein D. Alexander Graham Bell
Respuesta de asociación	Johannes Gutenberg — Teoría de la relatividad Thomas Alva Edison — El teléfono Albert Einstein — La imprenta Alexander Graham Bell — La bombilla
Respuesta de elección múltiple compleja	Indica cuales de las siguientes afirmaciones son ciertas: - La teoría de la relatividad la desarrolló Albert Einstein junto con Thomas Alva Edison. Verdadero/Falso - La bombilla la inventó Thomas Alva Edison. Verdadero/Falso - Thomas Alva Edison inventó el fonógrafo. Verdadero/Falso

Preguntas abiertas de respuesta abierta, construida o elaborada:

- Respuesta elaborada corta (precisan de correctores con pormenorizados criterios de valoración).
- Respuesta elaborada larga (precisan de correctores con pormenorizados criterios de valoración).
-

Pregunta de respuesta abierta

Respuesta elaborada corta	Cita dos inventos de Thomas Alva Edison: 1. _____ 2. _____
Respuesta elaborada larga	¿Qué supusieron los inventos de Thomas Alva Edison en su época? _____ _____

Las preguntas **no se presentan aisladas**, sino formando grupos bajo una presentación textual y/o gráfica común, que denominamos "**estímulo"**, que presenta al alumno una situación cotidiana como las que se pueden encontrar en la vida real.

Elaboración de ítems

La elaboración de **un ítem** y su inclusión dentro del conjunto de la prueba debe permitir la medida del rendimiento de los alumnos en la competencia evaluada, **debe discriminar**.

Crear un ítem puede ser relativamente sencillo. **Crear ítems** que se adapten a los descriptores establecidos en la matriz de especificaciones y además en un número elevado para poder hacer una selección posterior, según los porcentajes que se establezcan para la prueba, tanto de conocimientos como de procesos cognitivos, ya no **es tan sencillo**.

Además hay que crearlos vinculados a estímulos y esto supone un riesgo: en el caso de que se detecte que el estímulo no es adecuado, todos los ítems creados dependientes de él quedan anulados. Es por tanto necesario crear un número elevado y establecido desde el principio, adaptándose a las especificaciones de la matriz para evitar en la medida de lo posible trabajos innecesarios, o una mala calidad de los ítems y los estímulos.

Requisitos para la construcción de ítems

- Conocimiento en profundidad de los **contenidos** de la materia a evaluar.
- Conocimiento de los **descriptores** de cada dominio o proceso cognitivo.

Estos dos puntos son los que definen la matriz de especificaciones de una prueba y por tanto es el elemento de partida para comenzar la creación y construcción de preguntas de manera organizada. Al partir de una clasificación estructurada (descriptores) se evita producir un exceso o defecto de ítems para cada uno de los contenidos y dominios de la materia en la que se centra la prueba.

Un ítem que ha sido escrito considerando un determinado nivel cognitivo y un contenido concreto, tiene más probabilidades de ser un buen ítem.

- Cada prueba tiene una matriz de especificaciones concreta.
- La matriz es la herramienta que guía la construcción y la interpretación de la prueba.
- Además, es el instrumento de comienzo y es fundamental para el análisis final.
- Antes de elaborar los ítems es necesario tener la **matriz de la prueba** diseñada:
- Establecer la competencia.
- Determinar los contenidos.
- Determinar los descriptores.
- Determinar el peso relativo de cada casilla.
- Elaborar ítems (varios) específicos para cada descriptor.

Recomendaciones para la creación de un ítem

Recomendaciones para la redacción del tronco

- Puede ser una pregunta o una frase a completar.
- Ha de ser claro. Favorecer la claridad y la concisión.
- No debe utilizarse para enseñar, los ítems no tiene como fin la instrucción.
- Debe proporcionar sólo información precisa para responder.
- Emplear formas positivas al redactar los textos siempre que sea posible.
- Evitar las negaciones y, sobre todo, la doble negación.

Recomendaciones para las opciones (Clave y Distractores)

- La opción correcta (clave) debe ser claramente correcta. Si dos expertos albergan dudas acerca de la validez de la clave, habrá que revisar cómo se ha expresado o redactado el tronco y las opciones (alternativas), y analizar el porqué de la diferencia de criterio. Si no hay acuerdo, no es un buen ítem y no debe ir a una prueba.
- Los distractores que confunden a los expertos no son buenos.
- Podemos emplear preconceptos erróneos como distractores.
- La dificultad del ítem se puede variar haciendo los distractores similares a la clave.
- No abusar de opciones “todos los anteriores” – “ninguna de las anteriores”.
- Si algo se repite en todas las opciones, valorar eliminarlo o llevarlo al tronco.
- Todas las opciones deben ser de la misma longitud y complejidad, y de similar estructura sintáctica. La respuesta correcta no debe ser la más larga, ni la más corta, ni la única distinta.
- Con más opciones hay menor probabilidad de adivinación, pero también es más difícil la construcción.
- Es recomendable sólo una respuesta correcta, pues facilita el proceso de codificación.
- Evitar utilizar en la clave palabras que lleven a su selección por los que desconocen la solución. La repetición de palabras del tronco sólo en la clave es una pista que debe evitarse.
- Si el tronco es negativo, evitar opciones negativas.
- Distribuir la posición de la clave al azar. Evitar usar siempre la misma posición.

4. Pilotaje de la prueba objetiva

Las evaluaciones se construyen a partir de ítems y necesitamos que estos sean de buena calidad. Cualquier prueba puede ser mejorada a través de una adecuada selección de los elementos (ítems) que la componen. Una vez creados los ítems se hace necesario realizar un proceso de **pilotaje** que permita **analizar y calibrar** el trabajo realizado, **seleccionando** posteriormente aquellos que van a configurar la prueba.

El pilotaje y análisis de ítems es una parte muy importante en el proceso de construcción de un instrumento de este tipo.

Ya hemos dicho en repetidas ocasiones que un descriptor apunta a un bloque de contenidos y a un determinado proceso, por tanto el autor debe desarrollar para cada descriptor múltiples ítems, cada uno de ellos vinculados a diversos estímulos.

Se denomina **pilotaje** al proceso en que se configuran cuadernillos (trozos del instrumento completo) para llevar a campo los ítems, con la intención de medir el ítem, no el instrumento, ni los resultados de los que realizan la prueba. En realidad se generan muchos instrumentos distintos de manera que se incluyan todos los ítems y hay que garantizar que se dé respuesta a todos ellos un número de veces suficiente.

Mediante el pilotaje se pretende **comprobar que los ítems cumplen la función** para la que han sido escritos, y que el **grado de dificultad**, basado en los aciertos de los alumnos, confirman el proceso cognitivo para el cual el autor los concibió. El pilotaje se realiza con alumnos de otros cursos o con la cohorte de alumnos correspondiente, dado que responder al ítem no debe ser la dificultad, sino validar que discrimina, que no hay errores de redacción, que se interpretan correctamente, etc. En el proceso detectamos qué ítems son problemáticos por diversas causas.

Una vez realizado todo el proceso se deben **elegir aquellos ítems que configurarán la prueba**, eliminando los que se han comportado mal o de manera ineficiente y eligiendo entre los restantes, de forma que la prueba final se corresponda con la matriz de especificaciones que hemos diseñado. Por otro lado, aquellos que son adecuados y no van a formar parte de la prueba podrán ser utilizados en otras ocasiones.

Por tanto, puede suponer un problema que muchos ítems de un estímulo sean rechazados, o cuando a pesar de no ser rechazados se elimina el estímulo por considerarlo inadecuado. Cuando se selecciona un ítem, automáticamente queda seleccionado el estímulo, pero no todos los ítems que se diseñaron para el estímulo irán a la prueba. Es por esto que es necesario desarrollar un amplio número de ítems por estímulo y, a su vez, un amplio número de ítems por descriptor de manera que al final tengamos suficientes elementos de calidad para poder cumplimentar la matriz de especificaciones y configurar la prueba.

Análisis de ítems: comparabilidad, discriminación y validez

Uno de los objetivos de las evaluaciones internacionales es la **comparabilidad a lo largo del tiempo** y, por eso, un conjunto importante de los ítems se mantienen de una edición a otra. Además hay que tener garantías de que los ítems que sustituyen a los que se liberan cumplen la función de sus antecesores. Es importante tener un gran banco de ítems contestados por un número suficiente de alumnos de manera que podamos tener información válida de los mismos.

Diversos parámetros que nos permiten analizar las características técnicas de un ítem:

- Dificultad,
- discriminación u homogeneidad, y
- validez.

Todos los elementos que configuran la prueba han de contribuir a diferenciar, a **discriminar entre los participantes**. Por lo tanto, los elementos no tienen que ser ni muy fáciles, ya que los contestarían bien todos los alumnos, ni muy difíciles, ya que nadie sabría responder. La dificultad de los elementos tendrá que estar entre estos dos extremos, para que puedan discriminar adecuadamente.

Del mismo modo que la discriminación del ítem tiene que ver con la fiabilidad, la **validez** del ítem tiene que ver con la validez del instrumento. Por ejemplo, en los casos en los que medimos alguna variable de rendimiento, la variable de contraste o criterio externo pueden ser las notas o calificaciones académicas obtenidas por los estudiantes (lo que no es aplicable a las evaluaciones externas dado el carácter confidencial de las mismas).

Para **seleccionar un ítem** que se incorporará a la prueba tendremos en cuenta los siguientes puntos:

1. Correlación del ítem con un criterio externo (si lo hay).
2. Correlación del ítem con la puntuación total del test (consistencia interna).
3. Dificultad del ítem (grado de adecuación de la competencia).
4. Covarianza del ítem en el test.

5. Construcción de la prueba objetiva definitiva

Una **prueba objetiva o cuestionario de rendimiento** es un instrumento que recoge una serie de situaciones "problemáticas", preparadas y examinadas con anterioridad, a la que el alumno tiene que responder siguiendo unas instrucciones o reglas. Posteriormente se estiman sus respuestas comparándolas con las respuestas del grupo normativo y así se estima la calidad y el nivel. Para todo lo anterior necesitamos dos características importantes:

• Las **puntuaciones** tienen que **poder compararse** con las puntuaciones conseguidas por otros participantes.

• Las **condiciones** de aplicación, codificación (corrección) e interpretación han de ser las **mismas para todos**.

Cuando se planifica una prueba hay que definir la población objeto del estudio y la **finalidad**:

• Población objeto de la muestra (edad, nivel, etc.).

• Objetivo de la prueba (diagnóstico, evaluación, propuestas de mejora, etc.).

Posteriormente hay que determinar el **contenido** del instrumento, las preguntas o técnicas a incluir y el límite de tiempo y número de ítems que conformaran la situación de prueba. Sin olvidar definir objetivos, contenidos y competencias específicos, ponderaciones de los mismos y tipos específicos de ítems a incluir.

Podemos establecer un **proceso general para la construcción de una prueba objetiva**:

- Definición de la finalidad y el propósito de la prueba: Marco de la evaluación.
- Elaboración de la matriz de especificaciones.
- Construcción y validación de ítems.
- Selección de ítems y construcción de instrumentos piloto.
- Pilotaje de los instrumentos.
- Análisis y selección de los ítems definitivos.
- Elaboración de la escala de rendimiento.
- Descripción e interpretación de la escala.
- Análisis y explotación de resultados.

En muchas ocasiones se presupone que en preguntas de elección múltiple el **azar** juega un papel importante. Un alumno que responde al azar tendrá preguntas muy sencillas mal y otras de elevada complejidad bien, lo que nos indicaría que el alumno está contestando de manera aleatoria.

Fiabilidad y validez del cuestionario de rendimiento

Fiabilidad

Es un concepto relacionado con la estabilidad de la medida proporcionada por la prueba, con su consistencia y con la predictibilidad de la misma.

- Cuando a las mismas personas se les pasa la misma prueba en diferentes momentos, si no han cambiado, deberían obtener las mismas puntuaciones (o muy parecidas).

Fiabilidad como estabilidad temporal (correlación entre test – retest).

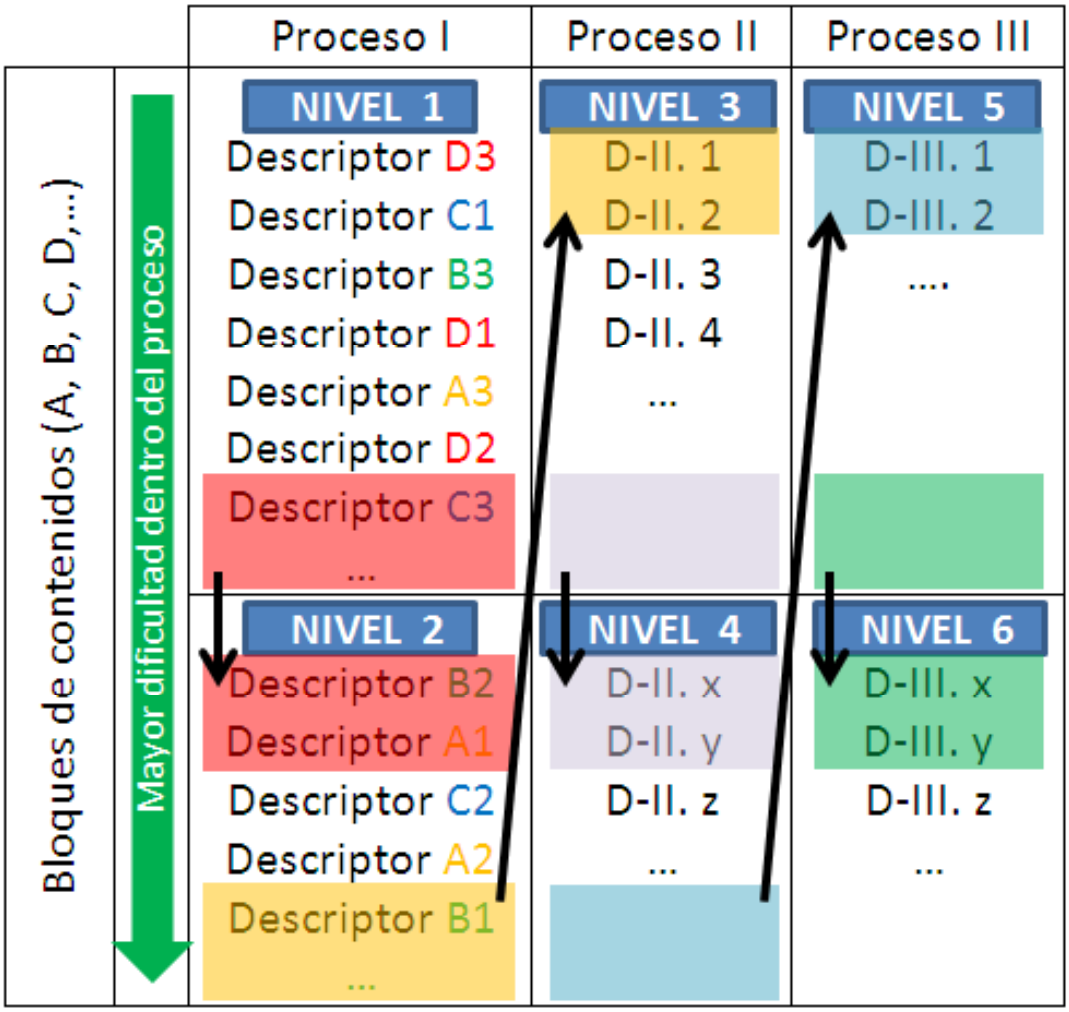
- De una prueba de competencia, por ejemplo en matemáticas, esperamos que todos sus elementos midan lo mismo (competencia matemática) y que, por tanto, sean sumables en una puntuación total única. Fiabilidad como consistencia interna.

Validez

- Grado en que mide lo que pretende medir y no otra cosa.
- Necesidad de referente o criterio externo (una prueba de matemáticas debe medir conocimientos matemáticos y no añadir dificultad por un uso complejo del idioma; una prueba de comprensión lectora no debe medir la corrección ortográfica del ejercicio).

¿Qué es una escala de rendimiento?

- Más allá de complejos análisis estadísticos, la escala de rendimiento la podemos entender como una **simplificación de la matriz de especificaciones** en la que se fijan unos niveles que definen lo que sabe hacer el alumno que es capaz de superar los ítems de ese nivel.
- Para ello, consideremos todos los **descriptores** de los bloques de contenidos asignados al proceso cognitivo I. Si reorganizamos todos los descriptores de mayor a menor facilidad, podemos hacer dos niveles de dificultad dentro del primer proceso cognitivo, dado que si, por ejemplo, hablamos de matemáticas habrá determinados descriptores de números más fáciles que otros de geometría, pero algunos de geometría serán más fáciles que alguno de números. Si repetimos el proceso con los demás procesos, podremos establecer seis niveles de rendimiento como, por ejemplo, establece PISA, o los cinco niveles que establece la IEA para PIRLS y TIMSS. En el siguiente esquema podemos ver gráficamente esta simplificación, aunque debemos destacar que PISA, al estar más alejada de los contenidos curriculares, incluye en los distintos niveles tareas de distinta consecución de la capacidad.



En la imagen se han definido cuatro bloques de contenidos (A, B, C y D) que se organizan de menor a mayor dificultad según los descriptores empleados para el proceso I. En el caso de los procesos II y III se han representado simplemente los de una organización de descriptores (D-II y D-III), sin especificar el bloque de contenidos al que pertenecerían.

En la imagen puedes observar cómo los descriptores de cada proceso se han clasificado según el nivel de dificultad, generando en el ejemplo seis niveles. Los recuadros marcados en rojo, naranja, morado, azul y verde son los ítems cuyos descriptores tienen la incertidumbre de permitir diferenciar entre niveles, discriminan el nivel en el que está el estudiante, confirmando con el resto de descriptores de ese nivel la pertenencia al mismo.

Tabla 3.3. Descripción de los niveles de Competencia matemática

Nivel	Lo que saben y lo que saben hacer los alumnos en cada uno de los niveles de rendimiento
5 646	En el *nivel 5* los alumnos además de los conocimientos y destrezas del nivel anterior, son capaces de: • resolver problemas relacionados con el entorno que exijan cierta planificación, aplicando dos operaciones con números naturales como máximo.
4 567	En el *nivel 4* los alumnos además de los conocimientos y destrezas del nivel anterior, son capaces de: • convertir unas unidades de medida en otras, • emplear fracciones usuales, como partes de la unidad, con denominador igual o menor de diez, • interpretar diferentes representaciones espaciales de objetos.
3 488	En el *nivel 3* los alumnos además de los conocimientos y destrezas del nivel anterior, son capaces de: • aplicar los conocimientos adquiridos sobre la medida, • apreciar si llegan a resultados válidos, exactos o estimados, en función de los números que intervienen y de la situación de cálculo en que se produce, • utilizar estrategias personales para la resolución de problemas y expresar de forma escrita y ordenada el proceso, • resolver problemas relacionados con el entorno que exijan cierta planificación, aplicando los contenidos básicos de geometría, • comunicar de forma escrita los resultados acompañados de una tabla, • utilizar las nociones básicas de los movimientos geométricos.
2 409	En el *nivel 2* los alumnos además de los conocimientos y destrezas del nivel anterior, son capaces de: • realizar cálculos con números naturales, usarlos en la resolución de problemas y dominar los algoritmos escritos, • utilizar técnicas sencillas de recuento, • recoger datos sobre hechos y objetos de la vida cotidiana, • utilizar las nociones básicas de los elementos geométricos (perímetro), • conocer las propiedades básicas de cuerpos y figuras planas, • utilizar los movimientos en el plano para emitir y recibir informaciones sobre situaciones cotidianas.
1 330	En el *nivel 1* los alumnos tienen capacidad para: • localizar y recuperar información sobre números, • interpretar el valor posicional de las cifras de un número, • utilizar las unidades de medida en situaciones cotidianas, • utilizar estrategias personales de cálculo mental, • realizar cálculos con números naturales y usarlos en la resolución de problemas, • expresar el resultado del recuento de datos en forma de tabla o gráfica, • interpretar un gráfico sencillo en una situación familiar y • reconocer y clasificar figuras y cuerpos geométricos.

6. Glosario de términos

(la definición de los términos que aparecen en este glosario han sido traducida y adaptada del glosario del Educational Testing Service, http://www.ets.org/understanding_testing/glossary/)

Análisis de ítems

Análisis estadístico sobre las respuestas de los examinandos a las preguntas de un test, hecho con el propósito de obtener información sobre la calidad de las preguntas de una prueba.

Banco de ítems

Base de datos para preguntas de pruebas (también su creación y mantenimiento). La grabación de cada pregunta incluye el texto de la pregunta e información estadística que se ha calculado a partir de las respuestas de los examinandos que las han realizado.

Calibración

El sentido de este término depende del contexto. En la Teoría de Respuesta al Ítem (TRI), la calibración se refiere al proceso de estimar los números (llamados parámetros) que describen las características estadísticas de cada pregunta del test. En la puntuación de un test de respuesta construida, la calibración se refiere al proceso de comprobar que con seguridad cada persona que puntúa aplica los criterios o estándares de corrección correctamente.

Capacidad (Habilidad)

El conocimiento, destrezas u otras características de los examinandos que se miden en una prueba.

Coeficiente alfa

Un estadístico que se usa para estimar la fiabilidad de las puntuaciones de un test. Lo que mide alfa es la consistencia interna – la cantidad en la que los examinandos ejecutan de forma similar los ítems. Bajo algunas concepciones que son generalmente razonables, alfa además indica hasta qué punto los examinandos realizarían la prueba de forma similar en dos modelos diferentes de la misma prueba. El coeficiente alfa se usa comúnmente para indicar la fiabilidad de las puntuaciones en las pruebas en las que todas las preguntas miden el mismo tipo de conocimiento o destreza. Sin embargo, puede usarse también para indicar el efecto halo entre las valoraciones que pretenden medir diferentes características de personas que están siendo evaluadas.

Coeficiente de fiabilidad

Un estadístico que indica la fiabilidad de las puntuaciones del test; es un estimador de la correlación entre las puntuaciones de unos examinandos en dos momentos de testaje con el mismo test (generalmente con diferentes modelos de la prueba).

Comparable

Dos puntuaciones son comparables si pueden compararse con sentido. Las puntuaciones directas de diferentes modelos de prueba no son comparables, porque las preguntas de un modelo pueden ser más difíciles que las de otro modelo. Las puntuaciones escaladas sobre diferentes modelos de un test son comparables si el proceso de calcularlas incluye la equiparación. Las puntuaciones percentiles son comparables si se refieren al mismo grupo de examinandos.

Correlación

Un estadístico que indica cuánto varían conjuntamente (covarían) dos medias, como puntuaciones de un test. Si la correlación entre las puntuaciones de dos tests es alta, los examinandos tienden a tener puntuaciones que están equiparadas por encima de la media (o por debajo) en ambos tests. La correlación puede variar desde –1 a +1. Cuando no hay tendencia en la covariación de las puntuaciones, la correlación es 0.

Correlación biserial

Estadístico usado para describir la relación entre la ejecución en un ítem y la de todo el test al que pertenece ese ítem. Es un estimador de la correlación entre la puntuación del test y una variable no observada que se asume para determinar la ejecución en un ítem, asumiendo además que tiene una distribución normal (la conocida curva de campana). Comparar con *correlación*, *correlación biserial-puntual*.

Correlación biserial-puntual

La correlación real entre una variable dicotómica (una variable con sólo dos posibles valores) y una variable con muchos valores posibles. Comparar con *correlación*, *correlación biserial*.

Descriptor del nivel de ejecución

Una declaración del conocimiento y destrezas que un examinando debe tener, para clasificarse en un nivel concreto de ejecución, como *básico*, *experto* o *avanzado*.

Desviación típica (de las puntuaciones de una prueba)

Una medida de la cantidad de variación en las puntuaciones de un grupo de examinandos. Es la distancia promedio de las puntuaciones respecto de la media (pero con la distancia promedio calculada mediante el procedimiento, raíz cuadrática de la media, que es algo más complicado que el método usual). La desviación estándar se expresa en las mismas unidades que las puntuaciones; p.ej.: número de respuestas correctas, o puntuación escala. Si hay muchas puntuaciones altas y bajas, la desviación estándar será elevada. Si las puntuaciones se concentran, la desviación estándar será pequeña.

Discriminación

Fuera del contexto de prueba, este término usualmente significa tratar a las personas de diferente forma por ser miembros de grupos particulares; p.ej.: hombre y mujer. En el contexto de prueba, la discriminación significa algo diferente. Se refiere al poder de una prueba o (de forma más extendida) una pregunta, para separar a los examinandos con mayor habilidad de los que tienen menos habilidad.

Diseño basado en la evidencia

Un enfoque para construir evaluaciones educativas que usan argumentos evidentes para revelar el razonamiento que subyace en el diseño de un test. Los diseñadores del test comienzan con un análisis de los tipos de evidencia necesarios para realizar afirmaciones válidas sobre lo que los examinandos pueden saber o hacer.

Distractores

En una pregunta de elección múltiple los distractores tienen respuestas incorrectas que se presentan a un examinando junto con la respuesta correcta. Los desarrolladores de preguntas usualmente usan distractores que representan errores comunes o desinformación.

Distribución normal

La distribución simétrica con forma de campana comúnmente usada en muchas aplicaciones estadísticas y de medición, especialmente en el cálculo de los intervalos de confianza incluyendo las bandas de puntuación.

Efecto halo

Cuando a los evaluadores se les pide que califiquen a personas sobre diversas cualidades diferentes, a veces tienden a calificar a cada persona de forma similar en todas estas cualidades, sin reconocer que algunas personas tienen puntuaciones más altas en unas habilidades que en otras. La tendencia de los calificadores a ignorar esta clase de diferencias se llama efecto halo.

Equiparación

Estadísticamente, ajustar las puntuaciones de diferentes modelos de la misma prueba para compensar las diferencias en dificultad. La equiparación hace posible informar sobre las puntuaciones escaladas que son comparables entre diferentes modelos de una prueba.

Error de clasificación

Ver *error de decisión*.

Error de decisión

Cuando las puntuaciones de los examinandos se comparan con un punto de corte especificado, son posibles dos tipos de errores de decisión: (1) un examinando cuya puntuación verdadera está por encima del punto de corte puede obtener una puntuación por debajo del punto de corte, (2) un examinando cuya puntuación verdadera está por debajo del punto de corte puede obtener una puntuación por encima del corte. Es posible modificar la regla de decisión para hacer que tipo de error ocurra con menos frecuencia, pero sólo con el coste de hacer el otro tipo de error de decisión más frecuente. Además denominado “error de clasificación”.

Error típico de medida (ETM)

Una medida que da la tendencia a variar de los examinandos debida a factores aleatorios, como la selección particular de los ítems de un modelo que se ha realizado, o algunos correctores que puntúan las respuestas de los examinandos. Cuanto más pequeño sea el ETM, menor será la influencia de estos factores. El ETM se expresa en las mismas unidades que las puntuaciones.

Escala de puntuación analítica

Un procedimiento para puntuar la respuesta sobre una prueba de respuesta construida, en la que la persona que puntúa proporciona puntos de acuerdo a rasgos específicos de la respuesta. Comparar con *escala de puntuación holística*.

Escala de puntuación holística

Un procedimiento para puntuar la respuesta de un test de respuesta construída, en el que el calificador realiza un único juicio sobre la calidad global de la respuesta, en lugar de proporcionar puntos de forma separada para diferentes rasgos de la respuesta. Comparar con *escala de puntuación analítica*.

Escalamiento

Transformar estadísticamente las puntuaciones de un conjunto de números (denominada "escala de puntuación") a otra. Algunos tipos de escalamiento se utilizan para hacer que las puntuaciones en diferentes pruebas sean comparables de alguna manera. La aplicación más común del escalamiento es hacer que las puntuaciones de diferentes ediciones ("modelos") del mismo test sean comparables. A veces los tests en diferentes materias se escalan para ser comparados en un grupo específico de examinandos. A veces pruebas de diferentes niveles de dificultad en la misma materia se escalan para que así estas puntuaciones escaladas en dos niveles adyacentes (p.ej: cursos) reflejen el mismo grado de habilidad en la materia; este tipo de escalamiento se denomina "escalamiento vertical".

Establecimiento de estándares

El proceso de seleccionar puntos de corte en una prueba.

Estaninas

Un tipo de puntuación referida a la norma, en el que sólo se dan puntuaciones de números enteros que van del 1 al 9. La escala de puntuación se define de tal manera que cada nivel de puntuación incluirá a un porcentaje específico del grupo normativo: pequeños porcentajes para los niveles más altos y bajos; grandes porcentajes para los niveles intermedios (Veáse *referencia a al norma*).

Evaluación; prueba; examen

Estos términos se refieren a todos los instrumentos o procedimientos para obtener información sobre el conocimiento, destrezas u otras características de las personas que son evaluadas, probadas o examinadas. Los tres términos se usan habitualmente de forma indistinta, pero hay algunas diferencias entre ellos. Evaluación es el más amplio de estos tres términos, examen el más específico.

Evaluación de ejecuciones

Un test en el que el examinando demuestra realmente las destrezas que la prueba pretende medir mediante la realización de tareas reales que requieren estas destrezas, en lugar de responder preguntas sobre cómo hacerlas. Generalmente, estas tareas requieren acciones muy diferentes a marcar un espacio en una hoja de respuestas o apretar un botón en un ordenador. Una prueba de lápiz y papel puede ser una evaluación de ejecuciones sólo si las destrezas que se miden se pueden mostrar, en un contexto real, con lápiz y papel. Comparar con *prueba de respuesta construida*.

Evaluación formativa

Evaluar las destrezas de los estudiantes con el propósito de planificar su instrucción. La evaluación formativa se realiza antes de que la instrucción comience y/o mientras está ocurriendo. Comparar con *evaluación sumativa*.

Evaluación no cognitiva

Intenta medir rasgos y conductas diferentes a las clases de conocimiento y destrezas medidas en las pruebas académicas más tradicionales – rasgos como "perseverancia, autoconfianza, autodisciplina, puntualidad, destrezas de comunicación, responsabilidad social y la habilidad para trabajar con otros y resolver conflictos" (R. Rothstein, The School Administrator, December, 2004; www.aasa.org/publications).

Evaluación sumativa

La evaluación de las destrezas de un estudiante con el propósito de determinar si la instrucción ha sido efectiva. La evaluación sumativa se realiza después de que la instrucción se ha completado. Comparar con *evaluación formativa*.

Fiabilidad

La tendencia de las puntuaciones de una prueba a ser consistentes en dos o más ocasiones de testaje, si no hubiese cambio en el conocimiento de los examinandos. Si un conjunto de puntuaciones tiene alta fiabilidad, las puntuaciones de los examinandos tenderían a coincidir mucho con sus puntuaciones en otro momento de testaje. Otro tipo de fiabilidad es la denominada *interjueces*, utilizada mucho en las puntuaciones que asignan diferentes puntuadores a un examinando en la realización de un ítem de respuesta construida.

Fórmula de corrección del azar

Una puntuación en la que cada respuesta incorrecta reduce la puntuación total del examinando en una fracción de punto. Esta fracción se elige para hacer que la ganancia esperada por una adivinación aleatoria se iguale a cero. Comparar con *puntuación de respuestas correctas*.

Funcionamiento Diferencial del Ítem (DIF: Differential item functioning)

Es la tendencia de una pregunta de una prueba a ser más difícil (o fácil) para ciertos grupos específicos de examinandos, después de controlar la habilidad de los grupos. Es posible realizar un análisis DIF para dos grupos de examinandos. En concreto un análisis DIF responde a "si comparamos estos dos grupos con el mismo nivel de habilidad global (indicado por sus ejecución en una prueba), ¿hay preguntas que son significativamente más difíciles para un grupo que para otro?.

Intervalo de confianza

El rango de valores posibles para una magnitud desconocida (como la puntuación verdadera de un examinando), calculado de manera que se tiene una probabilidad específica de incluir a dicha magnitud. Esta probabilidad específica se denomina "nivel de confianza" y es normalmente alta, generalmente 90 o 95.

Intervalo (rango, banda) de puntuación

Un intervalo alrededor de la puntuación de un examinando, proporciona la idea de que la puntuación individual de una persona está influenciada por factores aleatorios. Generalmente, los límites de la banda de puntuación están un error típico de medida por encima y por debajo de la puntuación real del examinando. Una puntuación determinada de esta forma es el intervalo de confianza a un nivel de confianza establecido, asumiendo una distribución normal, el 68%. Las bandas de puntuación ilustran la precisión limitada de la puntuación de una prueba como medida de algo más allá de la ejecución del examinando en más de una ocasión de prueba. Sin embargo, estas bandas pueden malinterpretarse de dos formas. Implican que la puntuación verdadera del examinando no puede estar fuera de este rango, e implican también que todos los valores en ese rango son valores igualmente probables para la puntuación verdadera del examinando. Ninguna de estas dos implicaciones es correcta.

Ítem

Una pregunta de un test, que incluye la pregunta en sí, cualquier material proporcionado como estímulo y las alternativas de respuesta (para un ítem de elección múltiple) o las reglas (escala de valoración) para un ítem de respuesta construida.

Ítem de elección múltiple

Una pregunta de una prueba que requiere que el examinando elija la respuesta correcta de entre un número limitado de posibilidades, generalmente cuatro o cinco. Comparar con *ítem de respuesta construida*.

Ítem de respuesta construida

Una pregunta que requiere que el examinando proporcione la respuesta en lugar de elegirla de una lista de posibilidades.

Ítem de respuesta cerrada

Cualquier tipo de ítem en el que la tarea del examinando es seleccionar la respuesta correcta de un conjunto de alternativas. Ítems de elección múltiple, ítems de verdadero-falso e ítems de asociación son todos ítems de respuesta cerrada. Comparar con *item de respuesta construida.*

Item dicotómico

Un item para el que hay solo dos posibles puntuaciones. Usualmente, 1 para la respuesta correcta y 0 para cualquier otra respuesta. Comparar con un *item politómico*.

Ítem politómico

Un ítem para que hay más de dos posibles puntuaciones (por ejemplo, un ítem con las puntuaciones siguientes: 0, 1, 2 y 3). Comparar con *ítem dicotómico*.

Media (de las puntuaciones de un test)

El promedio, calculado mediante la suma de las puntuaciones en los test de un grupo de examinandos, dividiendo por el número de examinandos en el grupo.

Mediana (de las puntuaciones de un test)

Es el punto en la escala de puntuación que separa a los examinandos en dos mitades. La mediana es el percentil 50.

Modelo de Rasch

Un tipo de modelo de la teoría de respuesta al ítem que asume que la probabilidad de un examinando de responder a una pregunta de un test depende únicamente de una característica de dicha pregunta, su dificultad. Comparar con *teoría de respuesta al ítem*.

Nivel de competencia

Declaraciones sobre el conocimiento, destrezas o capacidades de los examinandos que han conseguido un nivel específico de ejecución en la prueba. Los niveles de competencia pueden ser generales (por ejemplo, "El alumno/a puede leer en el 2º curso") o específicos (por ejemplo, "El alumno/a puede descodificar consonantes iniciales").

Normalización

Transformación de las puntuaciones de la prueba en una puntuación de escala de manera que produzca una distribución de puntuaciones que se aproxime a la distribución con forma de campana y simétrica denominada *distribución normal*. La normalización es un tipo de escalamiento.

Normas

Estadísticos que describen la ejecución de un grupo de examinandos (denominado *grupo normativo*) con el propósito de ayudar a los examinandos y examinadores a interpretar las puntuaciones. La información sobre normas se comunica generalmente en términos de rangos percentiles.

Parámetro *a*

En la teoría de respuesta al Ítem (TRI), el parámetro *a* es una magnitud que indica la discriminación de un ítem – con qué intensidad el ítem diferencia entre examinandos generalmente habilidosos y no habilidosos. Si el parámetro *a* es grande, la probabilidad de que el examinando contestará el ítem correctamente se incrementa mucho en un rango pequeño de habilidad. Si el parámetro *a* es pequeño, la probabildiad de responder correctamente se incrementa gradualmente en un amplio rango de habilidad.

Parámetro *c*

En la teoría de respuesta al ítem (TRI), el parámetro *c* es una magnitud que indica la probabilidad de que un examinando con poca o ninguna destreza o conocimiento de la materia contestará a la pregunta correctamente. Es decir, es el parámetro de adivinación.

Portfolio

Una recolección sistemática de materiales seleccionados para demostrar el nivel en conocimientos, destrezas o habilidades de una persona en un área en particular. Los portfolios pueden incluir documentos escritos (por la persona que va a ser evaluada u otras), fotos, dibujos, grabaciones de audio o video y otros medios. Generalmente los tipos de documento y otros medios proporcionados se especifican en detalle.

Prueba adaptativa

Un tipo de prueba en la que las cuestiones que se presentan al que realiza el test se seleccionan basándose en las respuestas previas que ha dado. Una buena ejecución en las anteriores preguntas lleva a que se presenten preguntas más difíciles y viceversa. El propósito de estas pruebas es el de hacer el tiempo de realización de una prueba más eficiente, no

proporcionando preguntas que son demasiado fáciles o demasiado difíciles para las personas que la realizan. Las pruebas adaptativas requieren procedimientos especiales para calcular las puntuaciones de los examinandos, ya que muchas combinaciones de preguntas diferentes son posibles, y algunos examinandos obtienen preguntas más difíciles que otros.

Prueba adaptativa computerizada

La prueba adaptativa se realiza con la ayuda de un ordenador. Por razones prácticas y logísticas, muchos tests adaptativos se proporcionan por ordenador.

Prueba de respuesta construida

Cualquier prueba en la que el examinando debe proporcionar la respuesta a cada cuestión, en lugar de elegirla de una lista de posibilidades. El término "prueba de respuesta construida" usualmente se refiere a una prueba que pide respuestas que pueden ser escritas en papel o tecleadas en un ordenador. Los tests que piden respuestas que no pueden ser escritas en papel o tecleadas se refieren generalmente a "pruebas de ejecución".

Psicómetra

Un experto en operaciones estadísticas asociadas con tests que miden características psicológicas, mentales, habilidades o de conocimientos y destrezas educativas u ocupacionales.

Puntuación baremada a nivel

Un tipo de puntuación normativa expresada en términos de ejecución típica de estudiantes de un nivel particular, en un punto particular del año académico. Por ejemplo, una puntuación de este estilo de 4,2 implica que la ejecución del examinando en el test sería típica para estudiantes en el segundo mes del cuarto nivel. (Ver *referencia a al norma*).

Puntuación de corte

Un punto en la escala de puntuaciones de la prueba utilizado para clasificar a los examinandos en grupos sobre la base de sus puntuaciones. A veces estas clasificacines se usan sólo para informar sobre estadísticas, como el porcentaje de estudiantes clasificados como habilidosos en una materia. Más habitual, las clasificaciones tienen consecuencias individuales para los examinandos – consecuencias como ser becado o denegar una licencia para practicar una profesión. (Ver además *descriptor del nivel de ejecución*).

Puntuación de respuestas correctas

Calcular la puntuación total contando el número de respuestas correctas, sin penalizar las respuestas incorrectas. Comparar con *fórmula de corrección por azar*.

Puntuación directa

Una puntuación de la prueba que no ha sido ajustada para poder ser comparada con las puntuaciones de otros modelos del test y no está expresada en términos de la ejecución de un grupo de examinandos. Los tipos más comunes de puntuaciones directas son el número de preguntas contestadas correctamente y, en un test de respuesta construida, la suma de las valoraciones asignadas por distintos evaluadores a las respuestas del examinando. Comparar con *Puntuación transformada*.

Puntuación objetiva

Un sistema de puntuación en el que una respuesta recibirá la misma puntuación, sin importar quien la puntúe. No se requiere de ningún juicio para aplicar la regla de puntuación. Comparar con *puntuación subjetiva*. Además, ver *escala de puntuación analítica* y *escala de puntuación holística*.

Puntuación percentil (rango percentil)

Una puntuación en la prueba que indica la posición relativa de un examinando en un grupo determinado. La puntuación percentil del examinando (también denominada *rango percentil*) es un número de 1 a 100 que indica el porcentaje del grupo con puntuaciones no mayores que la del examinando. La forma más común de calcular esta puntuación es calcular el porcentaje de personas del grupo con puntuaciones menores, más la mitad del porcentaje con exactamente las mismas puntuaciones que el examinando. (A veces ninguno de los examinandos con exactamente esa puntuación están incluidos, a veces todos ellos lo están). Las puntuaciones percentiles son fáciles de entender para la mayor parte de personas. Sin embargo, muchas personas no se dan cuenta que los promedios o diferencias de puntuaciones percentiles pueden confundir. Por ejemplo, la diferencia entre el percentil 90 y 95 casi siempre representa una mayor diferencia en rendimiento que la diferencia entre los percentiles 45 y 55. La comparación de los percentiles sólo tiene sentido si éstos se refieren al mismo grupo de personas que han realizado una prueba.

Puntuación subjetiva

Cualquier sistema de puntuación que requiera de un juicio por parte del evaluador (juez, puntuador). Con la puntuación subjetiva, diferentes puntuadores podrían asignar diferentes puntuaciones a una misma respuesta. Comparar con *puntuación objetiva*. Veáse además *escala de puntuación analítica* y *escala de puntuación holística*.

Puntuación transformada

Una puntuación del test que ha sido transformada en algo diferente a una puntuación directa. Un tipo común de puntuación transformada es la puntuación escalada – una puntuación que ha sido transformada en un diferente conjunto de números diferente al de las puntuaciones directas, generalmente después de equiparar para ajustar la dificultad de las preguntas de la prueba. Otro tipo de puntuación transformada son los percentiles. En lugar de puntuación transformada también se utiliza puntuación derivada.

Puntuación verdadera

En la teoría clásica, la puntuación verdadera de un examinando se define como el promedio de las puntuaciones que un examinando obtiene, promediado sobre un conjunto muy grande de condiciones posibles teóricas de testaje – por ejemplo, todas las posibles formas de un test, o todos los posibles correctores que puedan puntuar las respuestas. No es posible conocer la puntuación verdadera de un examinando, pero es posible estimar las puntuaciones verdaderas de un grupo de ellos.

Referencia al criterio

Hacer que las puntuaciones del test tengan sentido sin indicar la posición relativa de un examinando en un grupo. En un test de referencia al criterio, cada puntuación individual de un examinando se compara con un estándar fijo, en lugar de con la ejecución de otros examinandos. La referencia al criterio se define en términos de niveles de habilidad. La puntuación de la prueba requerida para lograr cada nivel de habilidad se especifica por adelantado. Los porcentajes de examinandos en cada uno de los niveles no son fijos, dependen de cómo lo hayan hecho los examinandos en la prueba Comparar con *referencia a la norma*.

Referencia a la norma

Hacer que las puntuaciones de la prueba tengan sentido proporcionando información sobre la ejecución de uno o más grupos de examinandos (llamados grupos normativos). Una puntuación referida a la norma indica la posición relativa del examinando en el grupo normativo. Un tipo común de puntuación referida a la norma son los percentiles. Otro tipo es la puntuación típica o estándar, que indica la posición relativa del examinando en términos de la media (puntuación promedio) y la desviación típica o estandar de las puntuaciones del grupo. Comparar con *referencia al criterio*.

Rúbrica

Conjunto de reglas para puntuar las respuestas de un ítem de respuesta construida. A veces denominada “guía de puntuación” o “escala de valoración”.

Teoría Clásica de los Tests (TCT)

Teoría estadística que forma la base de muchos cálculos realizados con puntuaciones de tests, especialmente los que tienen que ver con la fiabilidad. La teoría se basa en dividir la puntuación de una persona en dos componentes: un componente denominado “puntuación verdadera” que se generaliza a otras ocasiones de prueba con el mismo test, y un componente denominado “error de medida” que se estima utilizando el error estándar o típico de medida.

Teoría de Respuesa al Ítem (TRI)

Teoría estadística y el conjunto de métodos relacionados en los que la probabilidad de obtener cada una de las posibles puntuaciones en una pregunta del test depende de las características del examinando (llamada "habilidad") y un número pequeño (usualmente 3 ó menos) de características del ítem. Estas características del ítem o pregunta se indican con números llamados "parámetros". Siempre incluyen la dificultad de la pregunta y generalmente incluyen su discriminación.

Test de anclaje; ítems de anclaje

Para equiparar las puntuaciones de dos formas (modelos) de una prueba que son realizadas por diferentes grupos de examinandos es necesario saber cuánto difieren estos grupos en la competencia que mide la prueba. Un test de anclaje o prueba de anclaje es una prueba que se proporciona a ambos grupos para obtener esta información. La prueba de anclaje puede ser un conjunto de preguntas que aparecen en ambos modelos (ítems comunes o de anclaje), o puede ser una prueba separada que realizan ambos grupos.

Test estandarizado

Un test en el que el contenido y el formato así como las condiciones de prueba (tiempo, instrucciones, uso de calculadora) se controlan para que sea el mismo para todos los examinandos. (Se pueden especificar excepciones para examinandos con discapacidades).

Validez

Validez es el punto hasta el que las puntuaciones en una prueba son las apropiadas para un propósito específico. Las validez de las puntuaciones depende del modo en el que se interpretan y usan. Las puntuaciones en una prueba pueden ser válidas para un propósito determinado y nada válidas para otro. Las estadísticas pueden proporcionar evidencias para la validez de una prueba, pero la validez no puede ser medida mediante un único estadístico. Evidencias de validez pueden incluir:

- relaciones estadísticas de las puntuaciones de un test con otra información (p.ej.: puntuaciones en otras pruebas de la misma o similares habilidades, niveles educativos, juicios sobre el desempeño laboral)
- relaciones estadísticas entre partes de la prueba
- indicadores estadísticos de la calidad y equidad de las preguntas de la prueba
- la cualificación de los diseñadores, desarrolladores y revisores de la prueba
- el proceso de desarrollo de la prueba
- los juicios de expertos sobre hasta qué punto el contenidos de la prueba es coherente con el currículum o los requisitos de un trabajo (congruencia ítem-objetivo).