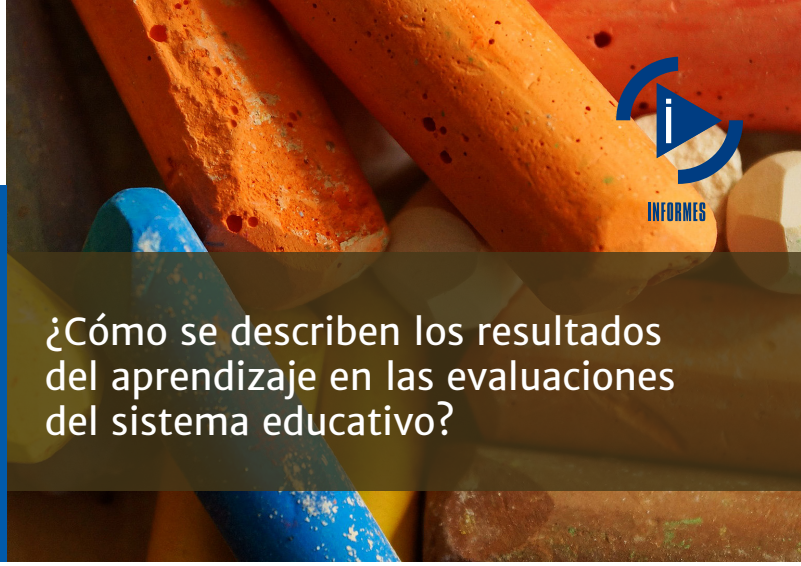
Gobierno del Principado de Asturias
Consejería de Educación y Cultura

Informes de Evaluación 14

Abril de 2018



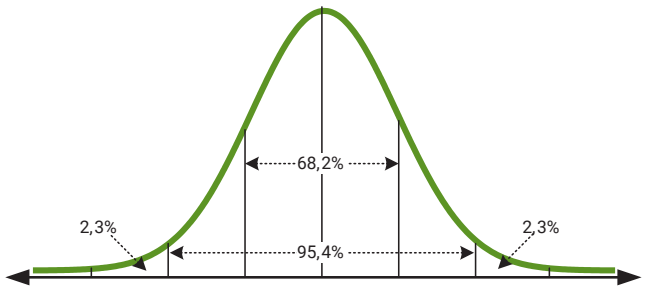
¿Cómo se describen los resultados del aprendizaje en las evaluaciones del sistema educativo?

Escalas de resultados y escalas de competencias: dos modos de expresar los logros del alumnado

Una de las finalidades primarias de la evaluación del sistema educativo es conocer y describir las competencias que adquiere el alumnado en el proceso de escolarización. Para ello se analizan sus respuestas a las pruebas cognitivas que evalúan las capacidades, destrezas y competencias de la población escolar. Las evaluaciones del sistema educativo expresan los logros del aprendizaje de dos maneras: escalas de puntuaciones o de resultados (del original en inglés *score scales* o *performance scales*) y escalas de competencia (del inglés, *proficiency scales*).

¿Por qué dos tipos de escalas?, ¿qué aportan las escalas de resultados y las escalas de competencias?

Las escalas de resultados *(score scales)* resumen el desempeño del alumnado mediante una puntuación numérica y continua (porcentajes de acierto, percentiles, puntuaciones típicas y transformadas, etc.). El gráfico 1 representa la *Escala N(500,100),* que es la puntuación transformada más empleada en la evaluación de sistemas educativos. Esta escala sigue la distribución normal y tiene de media 500 puntos y de desviación típica 100 puntos (de ahí su nombre).



Resultado en la escala N(500,100)	200	300	400	500	600	700	800
Resultado en puntos típicos (DT)	-3	-2	-1	0	+1	+2	+3
Porcentaje de población que queda por debajo de cada resultado	0,1%	2,3%	15,9%	50%	84,1%	97,7%	99,9%

Gráfico 1. Comparación de las puntuaciones de las escalas N(500,100), Puntos Típicos y Percentiles en la distribución normal

Establecer puntos de corte para dividir la escala de resultados es una tarea arbitraria pero muy práctica y eficiente

La escala *N(500,100)* tiene indudables ventajas: es sintética, estandariza los resultados de la población escolar (p. ej., se espera que aproximadamente el 68,2% de la población obtenga una puntuación comprendida entre 400 y 600 puntos) y permite comparar cualquier resultado con respecto al parámetro poblacional (p. ej., 600 puntos es un resultado satisfactorio ya que supera la puntuación del 84,1% de la población). Sin embargo, presenta una importante limitación, ya que no ofrece información sustantiva sobre los logros de aprendizaje y no responde a preguntas del tipo: ¿qué competencias tiene el alumnado que obtiene 600 puntos?, ¿qué sabe hacer el alumnado de una determinada edad?, ¿qué aprendizajes domina el alumnado con mayores dificultades de aprendizaje? Para responder a estas y otras cuestiones similares es necesario desarrollar escalas de competencia *(proficiency scales)* cuya función es traducir las puntuaciones numéricas de la escala de resultados a términos curriculares o de logro de aprendizajes.

¿Cómo se construyen las escalas de competencia ('proficiency scales')?

Construir escalas de competencia supone un esfuerzo colaborativo donde participan perfiles profesionales diversos (docentes y especialistas en el área evaluada, constructores de pruebas de evaluación y expertos en teoría del aprendizaje y de la medición educativa) en la realización de tres tareas o fases:

1. Determinar puntos de corte en la escala de resultados para establecer grupos de desempeño o niveles de rendimiento.
2. Asignar los ítems o tareas de las pruebas a los grupos o niveles de desempeño.
3. Elaborar descripciones que resuman las competencias del alumnado en cada uno de los niveles de desempeño.

1. Determinar puntos de corte y establecer grupos de desempeño

Establecer puntos de corte para dividir la escala de resultados en diferentes grupos o niveles de competencia es un procedimiento arbitrario pero, al tiempo, práctico y eficiente. Su lógica es muy similar al uso que la industria textil hace de las medidas antropométricas. Los rasgos físicos de las personas, igual que la métrica *N(500,100),* son muy variables y se expresan en escalas numéricas y continuas. Por ejemplo, el ancho de la cadera de los hombres oscila normalmente entre 65 y 150 centímetros, es decir, en un rango de 85 centímetros. Sin embargo, a efectos prácticos, la industria textil colapsa o agrupa este rango en unas pocas categorías: talla S (entre 78 y 85 cm.); talla M (entre 86 y 94 cm.); talla L (entre 95 y 99 cm.) y así sucesivamente. Salvando las distancias, determinar puntos de corte sigue la misma idea: arbitrar unos límites o intervalos en una escala continua para agrupar las puntuaciones en unos pocos niveles de desempeño.

Existen diferentes procedimientos para establecer puntos de corte. *La International Association for the Evaluation of Educational Achievement* (IEA) en sus evaluaciones de Compresión Lectora (PIRLS) y Matemáticas y Ciencias (TIMSS) señala cuatro puntos de corte en la escala *N(500,100):* 400, 475, 550 y 625 puntos, creando cuatro grupos con el alumnado que obtiene una puntuación de ±5 puntos sobre las marcas (Mullis, Cotter, Centurino, Fishbein, & Liu, 2016). Así, en el nivel Bajo agrupa el alumnado que logra entre 395 y 405 puntos en la prueba (gráfico 2).

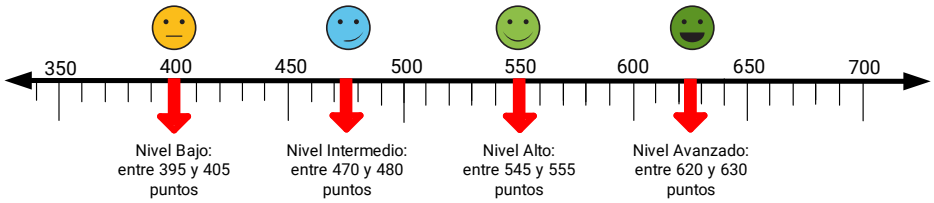


Gráfico 2. Puntos de corte en las evaluaciones PIRLS y TIMSS

Una variante de este procedimiento es fijar las marcas sobre la distribución de percentiles y no sobre puntos directos. La Evaluación de Diagnóstico de Asturias establece cinco puntos de corte en los percentiles 10, 25, 50, 75 y 90 y con ello se crean seis grupos o niveles de rendimiento (gráfico 3).

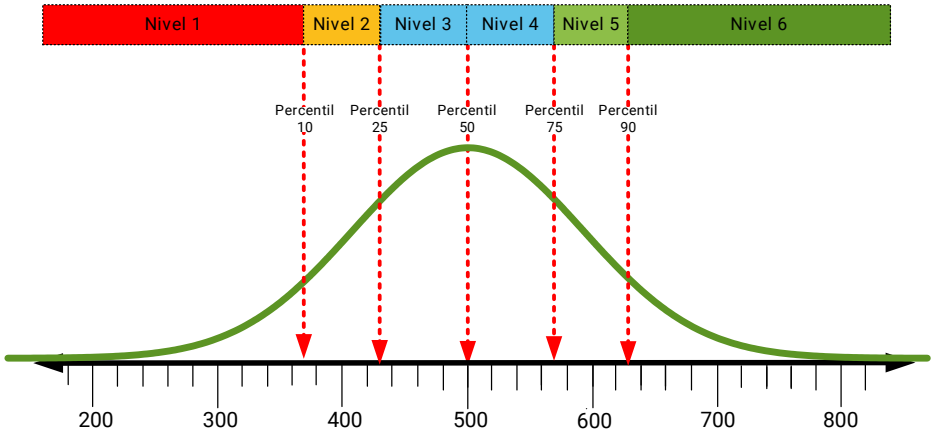


Gráfico 3. Puntos de corte en las Evaluaciones de Diagnóstico

En el *Programa Internacional para la Evaluación de Estudiantes* (PISA) el criterio para determinar los puntos de corte son las características de los ítems o preguntas de la evaluación y no un resultado numérico o un percentil establecido a priori (OECD, 2017). El gráfico 4 ejemplifica el procedimiento de PISA para establecer 6 grupos de rendimiento a partir de una prueba compuesta por 20 ítems.

En la parte central del gráfico se muestra conjuntamente la distribución de la escala de resultados que se ajusta a la curva normal (cada X representa 100 estudiantes) y los ítems ordenados por su dificultad. Disponer de la puntuación del alumnado y de la dificultad de los ítems en la misma escala de medida [*N(500,100)*] ofrece información relevante. Por ejemplo, señala que es muy probable que el alumnado con 600 puntos acierte los 15 ítems cuya dificultad está por debajo de esta marca (en este caso todos comprendidos entre el ítem 05 y el ítem 18), y también que será más probable que ese alumnado falle al responder los 5 ítems (ítems 09, 07, 20, 17 y 19) cuyo grado de dificultad está por encima de 600 puntos.

Para establecer los grupos de rendimiento el primer paso es decidir los puntos de corte inferior y superior de la escala.

- Determinar el límite inferior supone identificar inicialmente los ítems que preguntan por los aspectos básicos y elementales, de modo que el alumnado que no responda acertadamente a estos ítems se considerará poco competente, al menos en relación al resto de estudiantes que responden satisfactoriamente a los mismos. En este ejemplo, los ítems básicos fueron el 12, 13 y 18. El gráfico señala que el ítem 13 es el más difícil de los tres y, por tanto, el límite inferior de la escala, se

ubica inmediatamente por encima de la dificultad de dicho ítem, en este caso 360 puntos. Se dirá entonces que la probabilidad de acierto del alumnado que obtiene menos de 360 puntos en un ítem básico (ítem 13) es inferior al puro azar (p < 0,50).

- Para establecer el límite superior se opera del mismo modo. En este caso se determina inicialmente que sólo el alumnado más competente acertará los ítems 17 y 19. Entre ambos, el ítem 17 es más fácil y, por tanto, el límite superior se establece por debajo del nivel de dificultad de este ítem (680 puntos). De este modo se predice que un estudiante que obtenga más de 680 puntos tendrá una probabilidad superior al azar (p > 0,50) de acertar un ítem muy complejo. Por tanto, 680 puntos marca la diferencia entre el alumnado con mayor nivel de competencia y el resto del alumnado evaluado.

- Una vez se han determinado los puntos de corte superior e inferior el rango de puntuación comprendido entre ambos límites se divide en tantos niveles como sean necesarios en función del número de grupos. Dado que en este ejemplo se necesitan seis niveles de rendimiento, el espacio entre las marcas inferior y superior se divide en cuatro partes de 80 puntos cada una. Los ítems que se ubican dentro de cada cuadrante son asignados a su respectivo nivel de rendimiento.

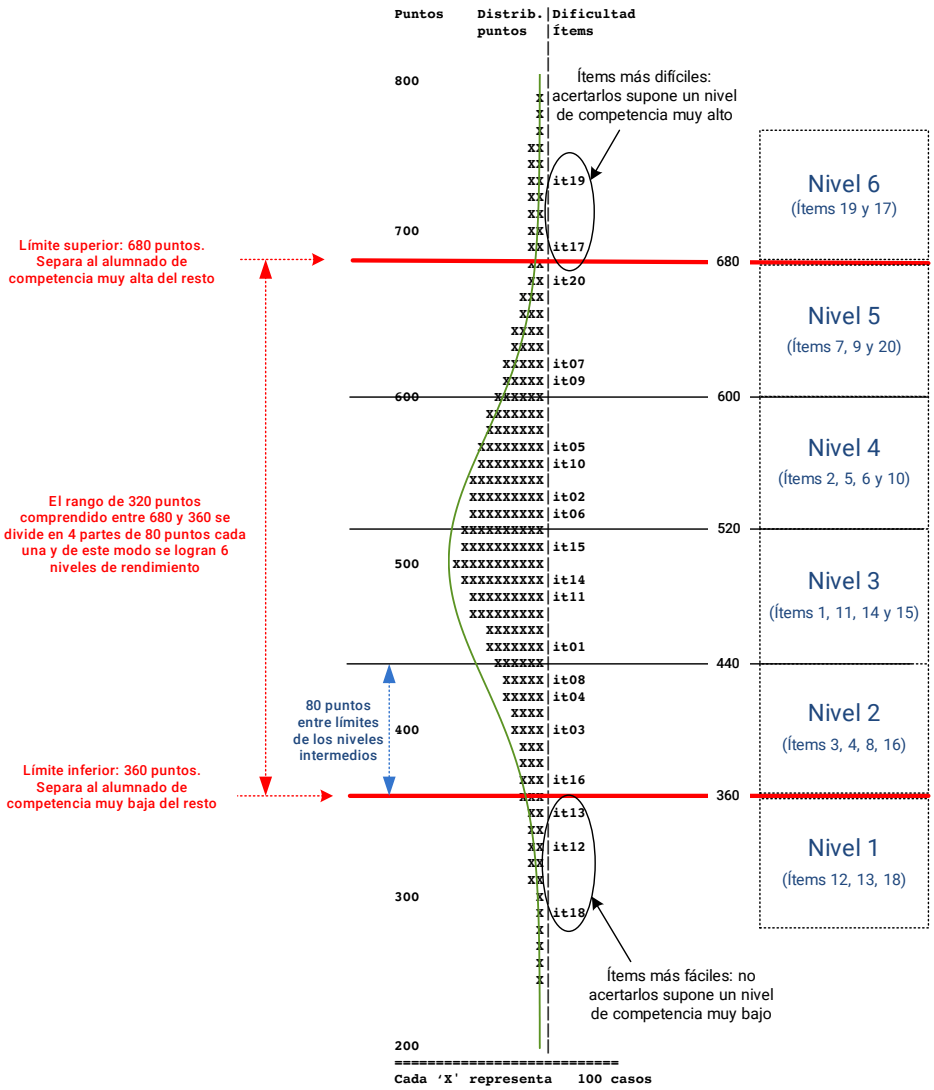


Gráfico 4. Puntos de corte en la evaluación PISA

2. Asignar ítems a niveles de rendimiento

Para asignar ítems a niveles de rendimiento se identifican los ítems que acierta la mayoría del alumnado de un determinado nivel de rendimiento y que al mismo tiempo falla la mayoría del alumnado del nivel inmediatamente inferior. Existen dos procedimientos según se comparen porcentajes o probabilidades de acierto.

TIMSS y PIRLS asignan ítems a niveles mediante la comparación de porcentajes. Como ya se vio (gráfico 2) estos programas crean cuatro grupos de rendimiento: Nivel 1 (Bajo, 400 puntos); Nivel 2 (Intermedio, 475 puntos); Nivel 3 (Alto, 550); y Nivel 4 (Avanzado, 625). A continuación se calcula el porcentaje de acierto de cada grupo en los ítems y éstos se asignan al nivel de rendimiento atendiendo a los siguientes criterios (Mullis et al., 2016):

- Ítems de Nivel 1 (Bajo). Un ítem se asigna al Nivel 1 si el alumnado del grupo Bajo presenta, como mínimo, un porcentaje de acierto del 65%. El conjunto de ítems de Nivel 1 describen las competencias del alumnado menos competente.

- Ítems de Nivel 2 (Intermedio). Un ítem será de Nivel 2 si el alumnado del grupo Intermedio presenta, como mínimo, un porcentaje de acierto del 65%, al tiempo que el porcentaje de acierto del grupo Bajo es inferior al 50%. El conjunto de ítems de Nivel 2 describe las competencias consolidadas por el alumnado del grupo Intermedio, pero que aún no domina el alumnado del grupo Bajo.

- Ítems de Nivel 3 (Alto). Un ítem será de Nivel 3 si el grupo Alto presenta, como mínimo, un porcentaje de acierto del 65% y el porcentaje de acierto del grupo Intermedio es inferior al 50%.

- Ítems de Nivel 4 (Avanzado). Un ítem será de Nivel 4 si el grupo Avanzado presenta, como mínimo, un porcentaje de acierto del 65% y el porcentaje de acierto del grupo Alto es inferior al 50%.

- Ítems difíciles para el Nivel 4. Un ítem será difícil para el Nivel 4 cuando el porcentaje de acierto del grupo Avanzado sea inferior al 50%. Estos ítems identificarían aspectos del programa de estudios que ni siquiera son dominados por el alumnado más competente.

El criterio 50-65% es bastante restrictivo. Así que, en ocasiones, con el fin de enriquecer las descripciones de los niveles de rendimiento los ítems también pueden anclarse o asignarse al nivel cuando el alumnado del nivel de rendimiento superior presente, como mínimo, un porcentaje de acierto del 60% a condición que el alumnado del grupo de nivel inferior no alcance el 50% de acierto.

El gráfico 5 muestra los porcentajes de acierto de cada grupo de nivel en cuatro ítems. El ítem 1 se asigna al Nivel 1 ya que el 70% del alumnado del grupo Bajo (N-1) responde al mismo satisfactoriamente. El ítem 2 es de nivel Intermedio (N-2), ya que la mayoría del alumnado de este nivel acierta el ítem, mientras que la mayoría del alumnado del grupo Bajo (el 60%) lo falla. El ítem 3 se asigna al grupo Alto (N-3), si bien en este caso se aplica el criterio más relajado del 60%. Finalmente el ítem 4 es demasiado difícil para el alumnado del nivel Avanzado ya que ni siquiera la mitad de los estudiantes de este grupo logran responder satisfactoriamente al ítem.

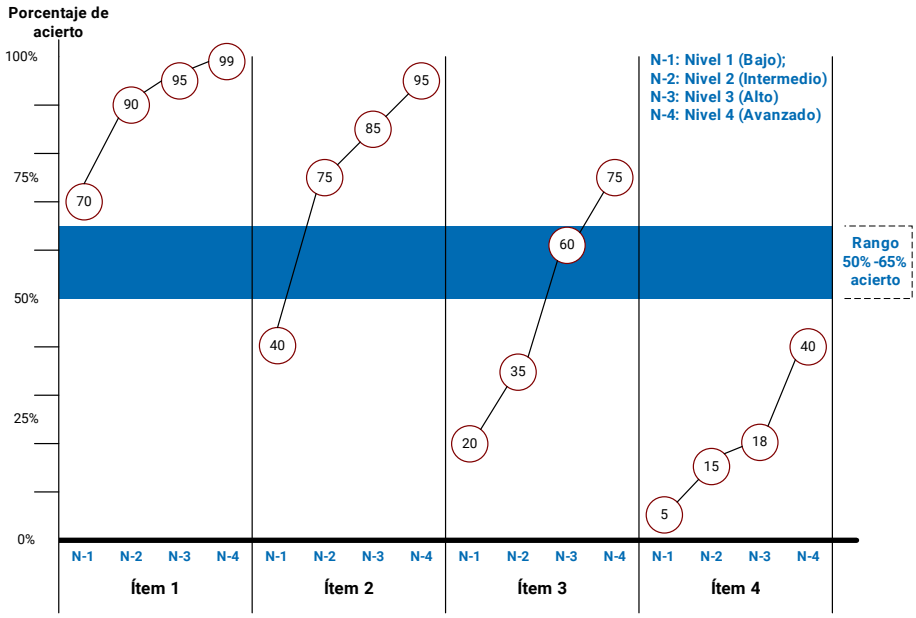
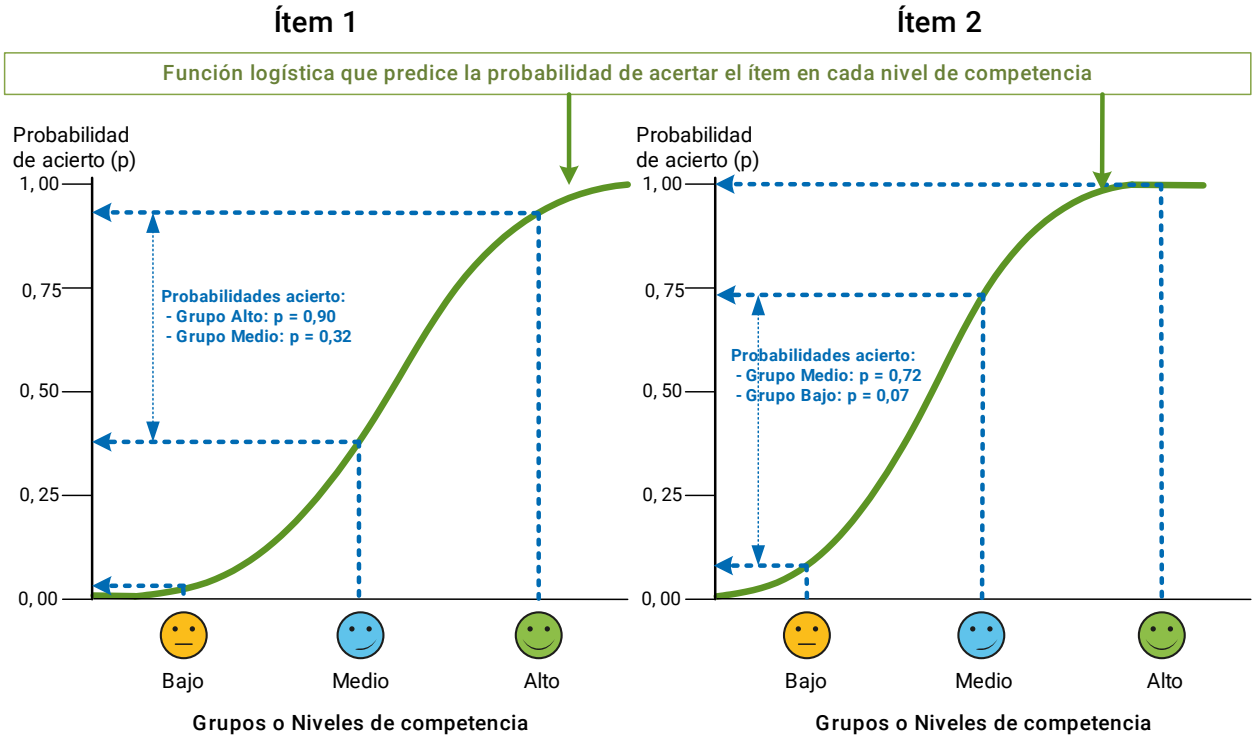


Gráfico 5. Porcentaje de acierto de los grupos o niveles de rendimiento en 4 ítems

Por su parte, PISA asigna los ítems a niveles de rendimiento mediante la comparación de probabilidades de acierto (gráfico 6). En este caso se han determinado tres grupos de rendimiento o competencia (Bajo, Medio y Alto). La función que predice la probabilidad de acertar el ítem 1 señala que es muy probable que, por su nivel de competencia, el alumnado del grupo Alto responda correctamente al mismo (p = 0,90), mientras que la probabilidad de acierto del alumnado del grupo Medio (p = 0,32) es inferior al acierto por azar (p = 0,50). Por tanto, el ítem 1 se asigna al grupo de competencia Alto, ya que solo el alumnado más competente parece tener probabilidades reales de acertar el ítem. Por su parte, el ítem 2 es un ítem del grupo de competencia Medio, ya que la probabilidad de acierto del alumnado de este grupo es 0,72, mientras que el grupo Bajo tiene una probabilidad de acierto muy baja.

Gráfico 6. Asignar ítems a niveles de rendimiento empleando las probabilidades de acierto



3. Elaborar y redactar descripciones que resuman los saberes, destrezas y competencias del alumnado en cada uno de los niveles de desempeño

Una vez que todos los ítems han sido distribuidos a su correspondiente nivel de desempeño se inicia la tercera fase de la tarea, que consiste en un análisis curricular de los ítems. Para ello se selecciona un panel de expertos en el área y curso evaluados que reciben los ítems ordenados por el nivel de desempeño. Las tareas a desarrollar por este grupo de expertos son las siguientes:

- Elaborar pequeñas descripciones de lo que supone responder correctamente a cada uno de los ítems. Generalmente estas descripciones son muy puntuales y concretas puesto que se refieren a los conocimientos y procesos cognitivos que se ponen en juego para responder satisfactoriamente el ítem. Ejemplos del tipo *"resolver operaciones de resta con llevadas en los supuestos más difíciles del algoritmo (inclusión de ceros en el minuendo y en el sustraendo)"*; o *"reconocer o identificar dos rasgos físicos del protagonista de una narración"*.

- El conjunto de ítems de cada nivel conforman el abanico de competencias, conocimientos y destrezas del alumnado de dicho nivel. Por ello, la segunda tarea consiste en redactar una descripción general que resuma y caracterice a cada uno de los niveles de desempeño. A continuación se ofrece una descripción general de la competencia matemática para el alumnado promedio de Asturias que finaliza Educación Primaria: *"El alumnado que se encuentra en el nivel intermedio es capaz de resolver problemas cotidianos reflexionando sobre el proceso seguido para solucionarlos, utilizando procedimientos matemáticos y algoritmos. Es capaz de identificar y usar las diferentes unidades del Sistema Métrico Decimal, de trabajar con escalas y de resolver problemas estadísticos sencillos. Utiliza las propiedades de las figuras planas e identifica y diferencia sus elementos"*.

> Las descripciones son elaboradas por paneles de expertos en el área y curso evaluados

- Seleccionar algunos ítems a liberar que ejemplifiquen las competencias propias de cada grupo o nivel de desempeño.

Características de las escalas de competencia ('proficiency scales')

Las escalas de competencia descritas tienen cuatro características:

- Son jerárquicas e inclusivas: se espera que el alumnado de un determinado nivel responda satisfactoriamente a los ítems de dicho nivel y a los ítems de niveles inferiores. Así, el alumnado del Nivel 2 acertará los ítems de su nivel y también los del Nivel 1.

- Las descripciones son probabilísticas, no deterministas. El alumnado de un determinado nivel tiene altas probabilidades de acertar los ítems de su nivel y de niveles inferiores, pero de ello no se sigue que el acierto esté garantizado. De hecho, las puntuaciones de dos estudiantes del mismo grupo o nivel de rendimiento pueden variar entre 75 y 100 puntos de la Escala *N(500,100)*. Por tanto, no puede concluirse que automáticamente todo el alumnado de un nivel responderá correctamente a los mismos ítems o que, por pertenecer al mismo nivel, su competencia real en la materia sea idéntica.

- Representan logros efectivos del alumnado. Las descripciones están extraídas del estudio empírico de los ítems de la prueba. Por tanto, sólo serán exhaustivas en la medida en que los ítems representen adecuadamente el contenido a evaluar.

- Tienen potencial para orientar la práctica educativa. Por la lógica de su construcción los niveles pueden predecir los próximos contenidos que dominará el alumnado. Así, en una escala de cuatro niveles, el alumnado del grupo Intermedio (Nivel 2) estará en condiciones de avanzar hacia los contenidos y destrezas del grupo Alto (Nivel 3), pero tendrá más problemas para dominar los del grupo Avanzado (Nivel 4). De alguna manera, los niveles de desempeño predicen los aprendizajes que está en condiciones de abordar con garantías el alumnado, convirtiéndose en una zona de desarrollo próximo.

En resumen...

- Las escalas de competencia *(proficiency scales)* traducen a términos curriculares y sustantivos los resultados numéricos alcanzados por el alumnado y expresan los logros del aprendizaje mediante descripciones detalladas de los conocimientos, habilidades y destrezas del alumnado.

- Construir escalas de competencia es un proceso en la que participan especialistas con diversos perfiles profesionales y que abarca tres grandes tareas: (1) establecer puntos de corte en la escala de resultados numéricos para crear grupos o niveles de rendimiento; (2) asignar los ítems empleados en la evaluación a cada nivel de rendimiento; y (3) elaborar descripciones que resuman los saberes, destrezas y competencias del alumnado en cada uno de los niveles de desempeño establecidos.

- Las escalas de resultados, pese a ser arbitrarias, presentan una serie de características: son jerárquicas e inclusivas; son probabilísticas y no deterministas, describen los logros efectivos del alumnado y tienen potencial para orientar la práctica educativa.

Referencias:

Mullis, I. V. S., Cotter, K. E., Centurino, V. A. S., Fishbein, B. G., & Liu, J. (2016). Using Scale Anchoring to Interpret the TIMSS 2015 Achievement Scales. En M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 14.1-14.47). Recuperado de: http://timss.bc.edu/publications/timss/2015-methods/chapter-14.html

OECD (2017). *PISA 2015 Technical Report.* Paris: OECD Publishing. Recuperado de: http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Edita: Consejería de Educación y Cultura del Gobierno del Principado de Asturias. Dirección General de Ordenación Académica e Innovación Educativa.
Autoría: Servicio de Evaluación Educativa.
D. Legal: AS 02537-2018