EDUCATIONIS MOMENTUM

Vol. 10, n.° 1, 2024, pp. 35-50, ISSN (online): 2517-9853

https://doi.org/10.36901/em.v10i1.1702

Una aproximación inicial a la optimización de la clasificación de estudiantes en evaluaciones de gran escala

An Initial Approach to Optimizing the Classification of Students in Large-Scale Assessments

Humberto Pérez León Ibañez

Universidad Católica San Pablo, Arequipa, Perú

hhperez@ucsp.edu.pe

https://orcid.org/0000-0002-6451-214X

Recibido: 2024.05.07

Aceptado: 2024.12.10

Resumen

La mayoría de los actos de evaluación implica clasificar a los estudiantes en categorías. Especialmente en las evaluaciones de altas consecuencias, esta clasificación conlleva una gran responsabilidad por las afectaciones que se derivan de ella. La puntuación bruta no refleja la naturaleza del error de medición y sus implicancias para la clasificación de los estudiantes. Por el contrario, el modelo Rasch ofrece un enfoque basado en la relación entre la probabilidad de acertar un ítem y la diferencia entre la habilidad del estudiante y la dificultad del ítem, aproximación que abre la posibilidad de calcular los errores estándar asociados a las medidas de ítems y personas. En el marco de este enfoque, este trabajo se propone investigar si la concentración de ítems alrededor de un punto de corte mejora la precisión de la clasificación, medida por el índice de precisión de Rudner. Se investigó esta relación en tres escenarios de dispersión de estudiantes. Los resultados muestran: a) que a mayor concentración de ítems alrededor del corte, mayor precisión de la clasificación; y b) que la fuerza de esta asociación se ve beneficiada en contextos de mayor concentración de la habilidad de los estudiantes.

Palabras clave: evaluaciones de gran escala, clasificación de estudiantes, modelo Rasch, precisión de la clasificación, error de medición

Abstract

Most evaluation activities involve classifying students into categories. Especially in high-stakes assessments, such classification carries significant responsibility due to the ensuing consequences. Raw scores do not reflect the nature of measurement error or its implications for student classification. In contrast, the Rasch model offers an approach based on the relationship between the probability of answering an item correctly and the difference between the student's ability and the item's difficulty, an approach that allows for the calculation of standard errors associated with item and person measures. Within this framework, this study aims to investigate whether concentrating items around a cutoff point improves classification accuracy, measured using Rudner's accuracy index. This relationship was examined under three scenarios with varying levels of student dispersion. The results show: (a) that the greater the concentration of items around the cutoff, the higher the classification accuracy; and (b) that the strength of this association is enhanced in contexts where students' abilities are more concentrated.

Keywords: large-scale assessments, student classification, Rasch model, classification accuracy, measurement error

Prácticamente todo acto evaluativo en el ámbito educacional implica también una clasificación o categorización del estudiante. La calificación vigesimal, vigente hasta hace algunos años, determinaba la nota 11 —independientemente de su significado o la ausencia del mismo— como la mínima aprobatoria. La llamada evaluación por competencias, que sugiere una mirada cualitativa del desempeño, también precisa de una clasificación en al menos tres categorías: en proceso, logrado y logro destacado. Los actos de clasificación son, pues, connaturales a los actos evaluativos.

Cabe precisar que no todos los actos clasificatorios tienen el mismo impacto en la vida de un estudiante. Algunos de ellos, los propios de la evaluación formativa, son, o deberían ser, usados para promover mejoras en la enseñanza y el aprendizaje en curso. En el argot de la evaluación se dice de ellos que son «de bajas consecuencias» o «de bajo riesgo» (*low stakes*). Otros actos clasificatorios son «de altas consecuencias» o «de alto riesgo» (*high stakes*), porque impactan de una manera decisiva en la vida del estudiante. Reprobar el año escolar o el examen de admisión a la universidad conllevan una carga emocional muy grande y pueden provocar enormes consecuencias futuras. En ese sentido, los actos clasificatorios en contextos de altas consecuencias implican una alta responsabilidad por parte del examinador, en tanto persona o en tanto sistema (Amrein & Berliner, 2002; Braun, 2004; Duffy et al., 2009).

Es frecuente dar por supuesto que las evaluaciones «objetivas», es decir, las que no requieren de juicio humano para su calificación, son intrínsecamente justas en el sentido en que la responsabilidad del logro depende únicamente del examinado (suponiendo que se han minimizado los errores en la construcción del instrumento, sesgos de algún tipo o barreras de acceso). Este es el punto que queremos problematizar en este artículo.

La clasificación en las evaluaciones de gran escala

Las evaluaciones de gran escala como las evaluaciones nacionales de logros de aprendizaje (ECE, EM), internacionales (PISA, ERCE) y exámenes de admisión a las universidades son ejemplos de evaluaciones que realizan clasificaciones a partir de las puntuaciones que los estudiantes obtienen en la resolución de pruebas. Se denominan evaluaciones referidas a una norma cuando la clasificación del estudiante se determina con relación a la distribución de un grupo de referencia (o grupo normativo) o alrededor de

un cálculo que puede involucrar la media, el rango y/o la distribución de la población. De cualquier forma, las evaluaciones referidas a normas dependen de las variaciones en el grupo de referencia. Por su parte, las evaluaciones *referidas a criterios* organizan la clasificación alrededor de estándares o puntos de corte en una escala, que permanecen estables en el tiempo, independientemente de los cambios en la distribución del constructo medido en la población (Shepard, 1979).

Imaginemos que deseamos conocer la verdadera habilidad de un estudiante en un determinado constructo. No nos bastaría con hacerle una pregunta; quizás necesitemos muchas más. En realidad, si queremos llegar al conocimiento exacto de su competencia, deberíamos ser capaces de administrarle todas las preguntas y tareas concebibles. Esto, por supuesto, es imposible, y por ello administramos solo un subconjunto de ellas: el test.

Un test constituye solo una muestra de una hipotética totalidad de preguntas y, por lo tanto, nuestro conocimiento de su verdadera competencia será incompleto. Este problema lo conocían ya los primeros investigadores en medición, de manera que la bien conocida Teoría Clásica de los Test (TCT) introdujo el concepto de «error» (Traub, 1997):

$$O = V + e$$

Donde:

O: puntaje observado en el test
V: puntaje verdadero del estudiante
e: error debido a ruidos en la medición

La TCT constituyó un importante avance en el establecimiento de un marco para la evaluación de la precisión de los test, desarrollando métodos como el alfa de Cronbach, ampliamente utilizado para medir la confiabilidad de las pruebas. No obstante, el problema de la clasificación implica la incertidumbre asociada a la medición del estudiante en relación con el estándar, problema que solo puede ser abordado indirectamente desde la TCT.

Desde un enfoque distinto, la Teoría de Respuesta al Ítem ha planteado nuevos y poderosos modelos que nos permiten estimar tanto el error estándar asociado a la dificultad de los ítems como el de la habilidad de las personas.

En el modelo de un parámetro logístico, o modelo Rasch (Hambleton et al., 1991; Molenaar, 1995; Yu, 2020), la probabilidad de acertar una respuesta frente a fallarla se expresa como una función de la diferencia entre la habilidad de la persona y la dificultad del ítem.

$$P\{X_{ni} = x\} = \frac{exp(\beta_n - \delta_i)}{1 + exp(\beta_n - \delta_i)}; x \in \{0,1\}$$

Donde:

X_{ni}: probabilidad del estudiante n de acertar el ítem i

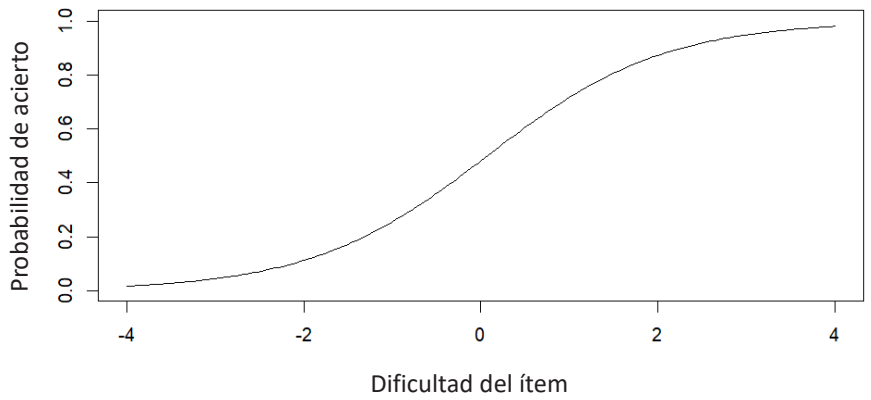
βn: habilidad del estudiante n

δi: dificultad del ítem i

Esta probabilidad se describe gráficamente mediante la curva característica del ítem, presentada en la figura 1.

Figura 1

Curva característica de un ítem



Nota. Elaboración propia.

Esta es una manera de ver la habilidad del estudiante (y de la dificultad de los ítems) en términos probabilísticos y no deterministas, y por tanto

radicalmente distinta a la planteada por la TCT. El modelo, entonces, establece que la distribución posterior de los parámetros de personas (e ítems) es una distribución normal con media = β_n y D.E. = E.E. El error estándar del modelo (Linacre, 2015) asociado a las medidas de las personas (y de los ítems) está dado por:

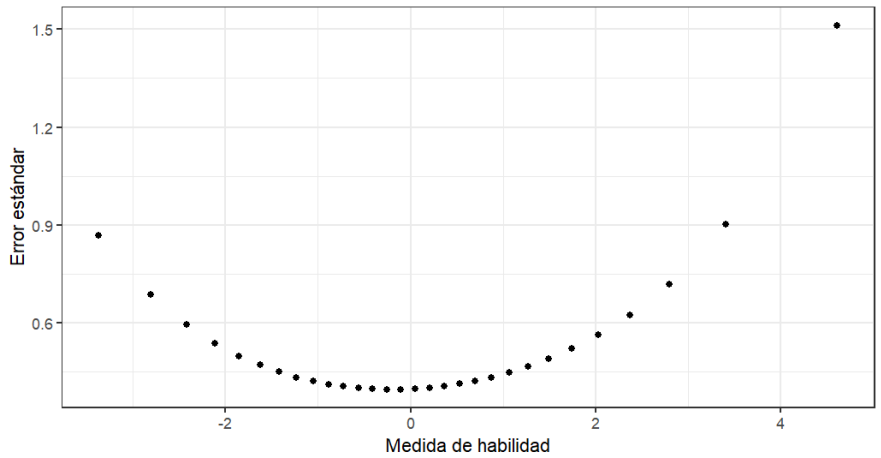
$$E.E.(\beta_n, \delta_i) = \frac{1}{\sqrt{\sum_1^N P_{ni}(1 - P_{ni})}}$$

Para efectos de ilustración, hemos simulado un conjunto de datos con la librería eRm (Mair & Hatzinger, 2007) y el análisis Rasch con la librería TAM (Robitzsch et al., 2024) del paquete estadístico R.

Como se puede observar en la figura 2, la distribución de los errores respecto de la habilidad de las personas es variable, y tiende a ser mayor en los extremos de la escala de habilidad, donde hay menos información, como es de esperar.

Figura 2

Error estándar versus medida de habilidad de las personas



Nota. Elaboración propia.

Ahora, supongamos un punto de corte arbitrario en la escala de la variable latente que sea la frontera para clasificar a los estudiantes en dos grupos, los que aprueban y los que no aprueban un test con altas consecuencias.

Teniendo el error estándar, estamos en condiciones de calcular intervalos de confianza para cada persona.

$$95\ \%\ IC_n = \beta_n \pm 1.96(E.E)$$

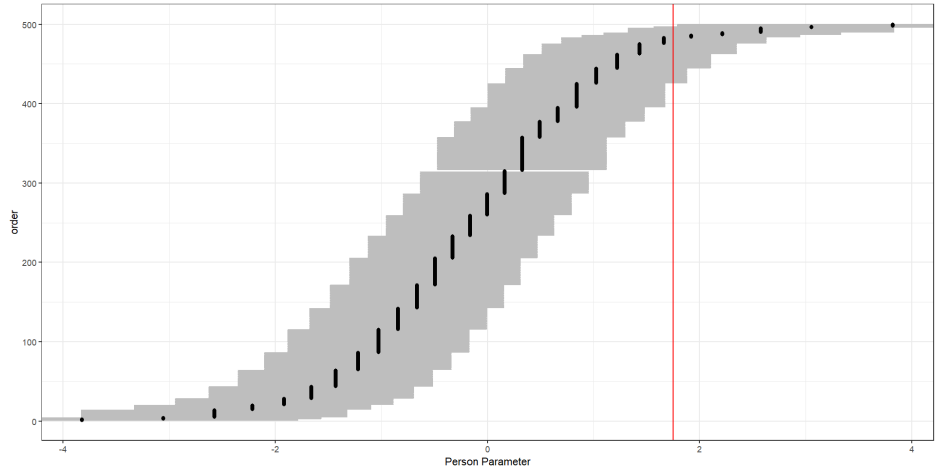
Lo cual se interpreta como que el verdadero valor de la habilidad de un estudiante se encontrará en 95 de 100 intervalos de confianza obtenidos al realizar la estimación múltiples veces.

Nos interesa aproximarnos a la incertidumbre de la clasificación hecha por un corte *q* en la escala de una variable latente debido al error estándar de medición. Entonces, empecemos por hallar cuál es la proporción de estudiantes en la muestra cuyos *IC* contienen el valor *q*.

En la figura 3 se muestra las medidas de las personas ordenadas por habilidad y sus correspondientes *IC* para un valor arbitrario de *q*.

Figura 3

Medida de las personas y sus intervalos de confianza al 95 %



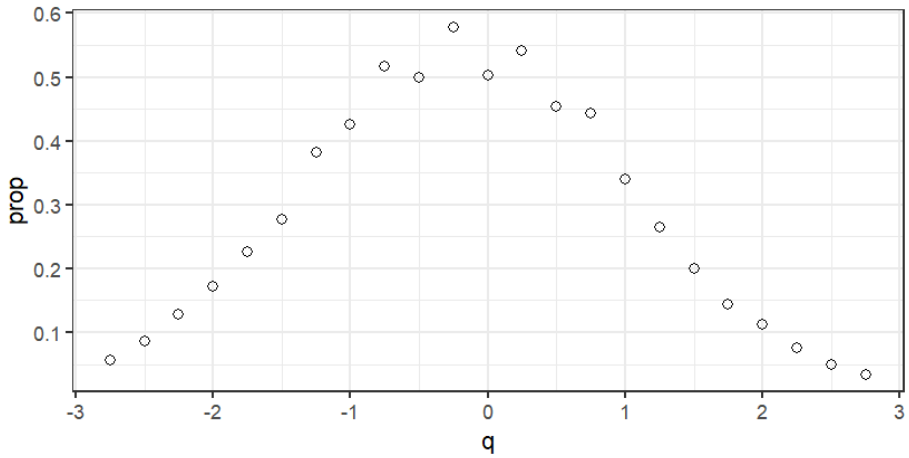
Nota. Elaboración propia.

Aproximadamente el 14 % de la muestra tiene un intervalo de confianza que contiene un valor arbitrario de $q = 1.75$. Si tomamos diversos valores de *q*, podemos observar que la proporción de la muestra que contiene este

valor en su *IC* varía notablemente (figura 4): es mayor en la parte media de la escala de habilidad y menor en los extremos. Esto es esperable, puesto que hay mayor proporción de la muestra concentrada en este sector.

Figura 4

Variación de proporción de estudiantes cuyo IC contiene valor q, para distintos valores de q



Una manera analítica de medir el error en la clasificación lo ofrece Rudner (2001, 2005). El objetivo de Rudner es, dado un punto de corte θ_c en la escala de la variable latente, identificar la proporción de falsos positivos y falsos negativos, tomando en consideración que la estimación puntual de habilidad de cada persona es la media de una distribución normal cuya desviación estándar es equivalente a su error estándar. Parte de la siguiente ecuación:

$$P_{(cm,n)} = \sum_{\theta_i < \theta_c}^{\theta_c} P(\hat{\theta} > \theta_c | \theta_i) f(\theta_i) \, / n$$

Donde $P(\hat{\theta} > \theta_c | \theta_i)$ representa la probabilidad de que el valor verdadero de habilidad del estudiante sea mayor que el valor del punto de corte (falso negativo), $f(\theta_i)$ es el número esperado de personas cuya medida de habilidad es θ_i y n es el total de examinados. Si consideramos a θ como una variable continua, entonces:

$$P_{(cm,n)} = \int_{-\infty}^{\theta_c} P(\hat{\theta} > \theta_c | \theta_i) f(\theta_i) / n$$

De la misma manera se obtienen los falsos negativos, así como los verdaderos positivos y negativos. A este indicador se le denomina índice de precisión de Rudner.

En este trabajo se desea evaluar si una mayor concentración de ítems alrededor de un punto de corte *q* mejora la precisión de la clasificación, medida por el índice de Rudner.

Método

Se simularon tres situaciones:

a. Primer escenario: cincuenta conjuntos de datos dicotómicos manteniendo la distribución de estudiantes constante con $N(0, 2)$, y estableciendo la distribución de dificultad de los ítems $N(1, DE)$, donde *DE* (desviación estándar) es un conjunto de cincuenta valores entre 0.1 y 4, uniformemente distribuidos.

b. Segundo escenario: cincuenta conjuntos de datos dicotómicos manteniendo la distribución de estudiantes constante con $N(0, 1)$, y estableciendo la distribución de dificultad de los ítems $N(1, SE)$, donde *DE* (desviación estándar) es un conjunto de cincuenta valores entre 0.1 y 4, uniformemente distribuidos.

c. Tercer escenario: cincuenta conjuntos de datos dicotómicos manteniendo la distribución de estudiantes constante con $N(0, 0.5)$,y estableciendo la distribución de dificultad de los ítems $N(1, DE)$, donde *DE* (desviación estándar) es un conjunto de cincuenta valores entre 0.1 y 4, uniformemente distribuidos.

Los conjuntos de datos fueron simulados con la librería eRm. Se eliminaron los ítems con respuestas perfectas o nulas en los conjuntos de datos donde aparecían, para evitar problemas en la estimación.

Posteriormente, en cada escenario se realizó el análisis Rasch de los cincuenta conjuntos de datos, mediante el algoritmo de máxima verosimilitud marginal (MML) y se obtuvo los estimados de habilidad de las personas y sus errores estándar mediante el método de estimación por verosimilitud ponderada (WLE). Estos cálculos fueron realizados mediante la librería TAM.

Por último, se calculó el índice de precisión de Rudner para cada resultado, haciendo uso de la librería cacIRT (Lathrop, 2015).

Resultados

De acuerdo con el objetivo de esta investigación, se desea conocer si una mayor concentración de ítems alrededor del punto de corte *q* se traduce en una mayor precisión de la medición de la habilidad. En este caso, se ha elegido el punto q = 1 como punto de corte arbitrario y se ha hecho variar la dispersión de los ítems alrededor de ese punto de corte, y de las personas alrededor de su media de habilidad.

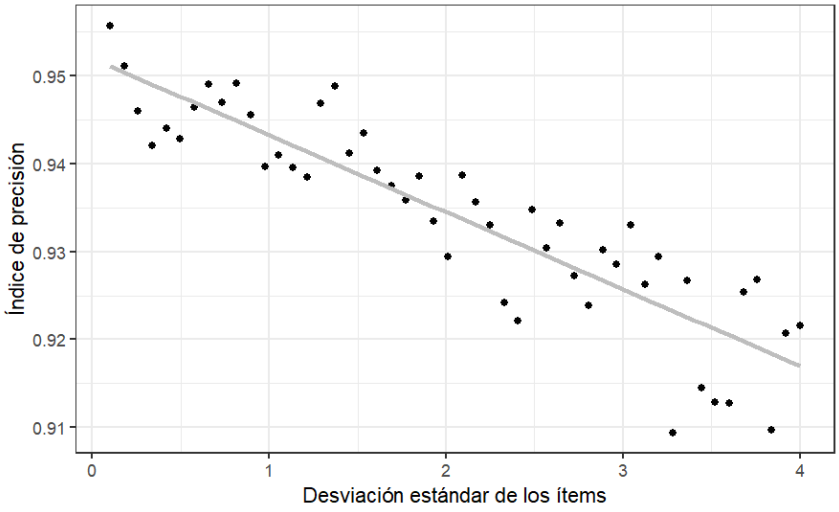
Una vez obtenidas las medidas de precisión de Rudner, se procedió a establecer si había una relación entre la desviación estándar de los ítems y dicho índice en cada uno de los tres escenarios.

Escenario 1. Personas con N(0, 2)

La relación se aprecia gráficamente en la figura 5.

Figura 5

Relación entre desviación estándar de ítems alrededor del punto de corte y el índice de precisión de la clasificación



Se llevó a cabo una regresión lineal simple, cuyos resultados fueron: *F (188.6, 48), p < 0.05, R2 = 0.80*.

Tabla1

Regresión lineal simple para el escenario 1

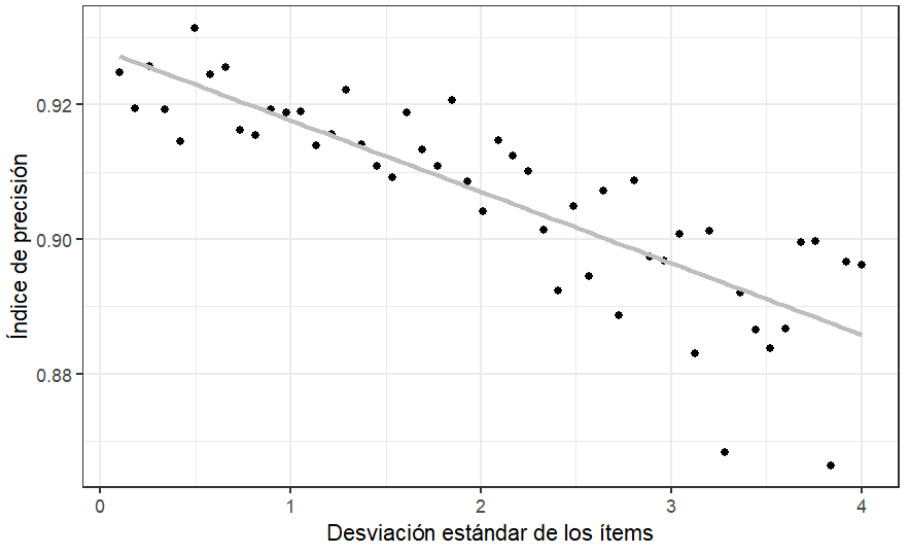
	Beta	E.E.	p
Intercepto	0.95	0.001	<0.001
D.E.	-0.009	0.001	<0.001

Escenario 2. Personas con N(0, 1)

En la figura 6 se muestra la variación de la precisión de la clasificación con relación a la desviación estándar de los ítems en el escenario 2.

Figura 6

Relación entre desviación estándar de ítems alrededor del punto de corte y el índice de precisión de la clasificación



Se llevó a cabo una regresión lineal simple, cuyos resultados fueron: *F (118.6, 48), p < 0.05, R2 = 0.71*.

Tabla 2

Regresión lineal simple para el escenario 2

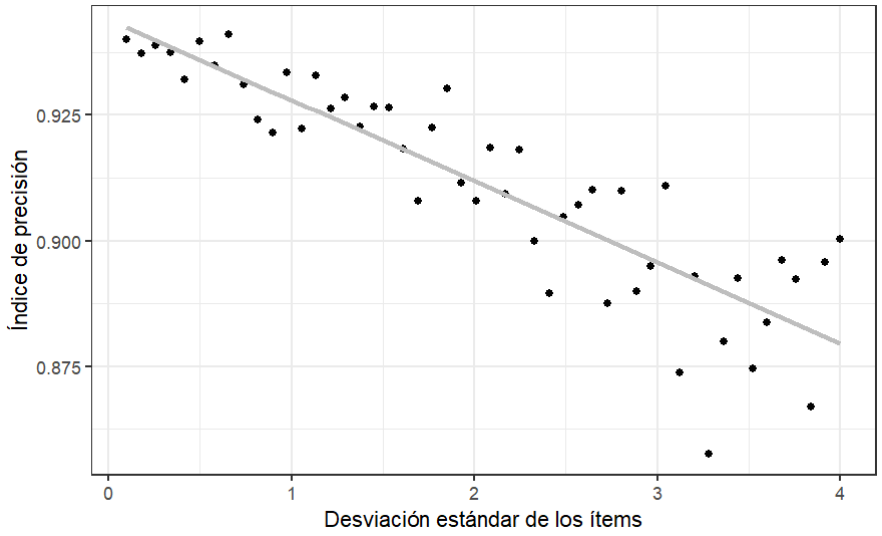
	Beta	E.E.	p
Intercepto	0.93	0.002	<0.001
D.E.	-0.011	0.001	<0.001

Escenario 3. Personas con N(0, 0.5)

En la figura 7 se puede apreciar la relación entre el índice de precisión de clasificación y la desviación estándar de los ítems para el escenario 3.

Figura 7

La relación entre el índice de precisión de clasificación y la desviación estándar



En este caso también se llevó a cabo una regresión lineal simple, cuyos resultados fueron: $F\ (168.6,\ 48),\ p < 0.05,\ R2 = 0.78$.

Tabla 3

Regresión lineal simple para el escenario 3

	Beta	E.E.	p
Intercepto	0.94	0.002	<0.001
D.E.	-0.016	0.001	<0.001

Como se puede apreciar, en los tres escenarios hay una dependencia negativa de la precisión de la clasificación con la dispersión de los ítems alrededor del punto de corte *q*. Así mismo, se puede observar que la fuerza de esa dependencia negativa va incrementándose a medida que se incrementa la concentración de los estudiantes.

Discusión

Se ha señalado que los actos evaluativos, por lo general, tienden a clasificar a los estudiantes en categorías. En muchos casos esta clasificación tiene importantes consecuencias para la vida de los estudiantes, como en los exámenes de admisión.

Desde un punto de vista determinista, la puntuación total del estudiante en la prueba es suficiente para determinar si el estudiante supera o no una valla o un estándar. Sin embargo, esta manera de ver el rendimiento no tiene en consideración que la estimación de la habilidad de una persona siempre está sujeta a un error de medición que puede ser estimado.

En este trabajo se investigó cuáles eran las consecuencias de manipular las condiciones de concentración de la dificultad de los ítems y habilidad de los estudiantes, dado un punto de corte q.

Los resultados muestran que en los tres escenarios existe una asociación negativa entre la dispersión de los ítems alrededor del punto q y el índice de precisión de Rudner. Es decir, que a mayor concentración (menor desviación estándar) de los ítems mayor será la precisión de la clasificación. Además, se mostró que la concentración de las personas también influye en la fuerza de esta asociación. Esta aumenta conforme aumenta la concentración de los estudiantes.

Las implicancias de estos hallazgos en el diseño de los test tienen que ver con la estrategia de acumular ítems alrededor del punto de corte para optimizar la clasificación, incrementando con esto la cantidad de información y disminuyendo los errores estándar. Esta estrategia se vería beneficiada en escenarios donde la medida de habilidad de las personas muestra mayor concentración.

En entregas posteriores sobre este tema se tiene planeado observar de forma sistemática la influencia de distintas distribuciones sobre la consistencia y precisión de los ítems, y distintas ubicaciones del punto de corte q, entre otras variables.

Por último, se postula que aplicar el enfoque de la Teoría de Respuesta al Ítem en evaluaciones de altas consecuencias puede ayudar a mejorar la justicia en la clasificación de las personas.

Referencias

Amrein, A. L., & Berliner, D. C. (2002). High-Stakes Testing & Student Learning. *Education Policy Analysis Archives*, *10*, 18. https://doi.org/10.14507/epaa.v10n18.2002

Braun, H. (2004). Reconsidering the Impact of High-Stakes Testing. *Education Policy Analysis Archives*, *12*, 1. https://doi.org/10.14507/epaa.v12n1.2004

Duffy, M., Giordano, V. A., Farrell, J. B., Paneque, O. M., & Crump, G. B. (2009). No Child Left Behind: Values and Research Issues in High-Stakes Assessments. *Counseling and Values*, *53*(1), 53-66. https://doi.org/10.1002/j.2161-007X.2009.tb00113.x

Hambleton, R. K., Swaminathan, H., & Jane Rogers, H. (1991). *Fundamentals of Item Response Theory*. SAGE.

Lathrop, Q. N. (2015). Practical Issues in Estimating Classification Accuracy and Consistency with R Package cacIRT. *Practical Assessment, Research & Evaluation*, *20*(18), 1-5.

Linacre, J. M. (2015). *WINSTEPS* (3.91.2.0).

Mair, P., & Hatzinger, R. (2007). Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software*, *20*(9). https://doi.org/10.18637/jss.v020.i09

Molenaar, I. W. (1995). Some Background for Item Response Theory and the Rasch Model. En I. W. Fischer Gerhard H. & Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (pp. 3-14). Springer New York. https://doi.org/10.1007/978-1-4612-4230-7_1

Robitzsch, A., Kiefer, T., & Wu, M. (2024). *TAM: Test Analysis Modules*. https://CRAN.R-project.org/package=TAM

Rudner, L. M. (2001). Computing the Expected Proportions of Misclassified Examinees. *Practical Assessment, Research & Evaluation*, *7*(14), 1-5. http://pareonline.net/misclass/class.asp.

Rudner, L. M. (2005). Expected Classification Accuracy. *Practical Assessment, Research, and Evaluation*, *10*(13), 13. https://doi.org/10.7275/56a5-6b14

Shepard, L. (1979). Norm-referenced vs. criterion-referenced tests. *Educational Horizons*, *58*(1), 26-32.

Traub, R. E. (1997). Classical Test Theory in Historical Perspective. *Educational Measurement: Issues and Practice*, *16*(4), 8-14. https://doi.org/10.1111/j.1745-3992.1997.tb00603.x

Yu, C. H. (2020). Objective Measurement: How Rasch Modeling Can Simplify and Enhance Your Assessment. En M. S. Khine (Ed.), *Rasch Measurement: Applications in Quantitative Educational Research* (pp. 47-73). Springer Singapore. https://doi.org/10.1007/978-981-15-1800-3_4