INSTRUMENTOS DE EVALUACIÓN. INFLUENCIA DE FACTORES INHERENTES Y EXTERNOS A LOS TESTS OBJETIVOS SOBRE LA CALIFICACIÓN OBTENIDA POR LOS ESTUDIANTES EN SU APLICACIÓN

EVALUATION INSTRUMENTS. INFLUENCE OF INHERENT AND EXTERNAL FACTORS TO OBJECTIVE TESTS ON THE QUALIFICATION OBTAINED BY THE STUDENTS IN YOUR APPLICATION

Gracenea, Mercedes
Universidad de Barcelona
Instituto de Ciencias de la Educación
Departamento de Biología, Sanidad y Medioambiente
Facultad de Farmacia y Ciencias de la Alimentación
Avda. Joan XXIII, s/n, 08020 Barcelona, España
gracenea@ub.edu

Monleón Getino, Antonio
Universidad de Barcelona
Departamento de Genética, Microbiología y Estadística
Facultad de Biología
Avda. Diagonal, s/n, 08020 Barcelona, España
amonleong@ub.edu

1. RESUMEN

Los tests objetivos con respuesta verdadero/falso son instrumentos de evaluación controvertidos. Se analizan dos permutas de un test de 50 preguntas aplicadas a 358 estudiantes (cinco grupos), en la evaluación continuada de una asignatura de grado. Se determina la posible influencia de la permuta, grupo de clase y sexo sobre las calificaciones de los estudiantes mediante un modelo lineal de evaluación de la varianza. Permuta y sexo no influyen significativamente pero sí el grupo clase.

2. ABSTRACT

Objective tests with true/false answers are controversial assessment instruments. Two swaps of a 50-question test are applied to 358 students (five groups), in the continuous evaluation of a subject of degree. The possible influence of the swap, as inherent factor to the test, and of class group and sex, as external ones, on students' grades is determined through a linear model of variance evaluation. Swap and sex do not significantly influence but the class group does.

3. PALABRAS CLAVE

Evaluación, test objetivo, índice de dificultad, índice de discriminación

4. KEYWORDS

Avaluation, objetive tests, dificulty index, dicrimination index

5. DESARROLLO

INTRODUCCIÓN

Los procesos de evaluación empleados en las asignaturas del ámbito universitario diseñan un escenario específico en el marco del cual los estudiantes reciben información sobre el nivel competencial que adquieren en las áreas de docencia-aprendizaje definidas por dichas asignaturas. Los instrumentos empleados en la evaluación, tanto si es entendida en su aspecto más clásico (Tyler, 1950, 1967, 1969), como si se considera bajo el prisma de investigación evaluativa (Joint Comittee, 1988; Mateo et al., 1993; Rodríguez Neira et al., 1995), o bien como proceso colaborativo de enseñanza-aprendizaje (Guba y Lincoln, 1982) o como evaluación de procesos contextualizados (modelo CIPP) (Stufflebeam, 2001), han de proporcionar resultados fiables, válidos y objetivos (Joint Comitee, 1988; Weir, 2005). En este contexto, los tests objetivos de evaluación con respuesta cerrada verdadero/falso constituyen una herramienta que puede resultar adecuada, fundamentalmente si la evaluación docente ha de ser realizada en grupos numerosos. Los aspectos positivos y negativos de estos instrumentos de evaluación han sido ampliamente tratados (Burton, 2004, 2005; Tasdemir, 2010), en sus diferentes dimensiones, ofreciendo características interesantes en su constructo relacionadas con la fiabilidad, la validez, la oportunidad o no de las puntuaciones negativas, e incluso valorando críticamente algunos mitos y malas interpretaciones. Los tests objetivos cerrados verdadero/falso pueden favorecer la memoria consciente y la retención a largo plazo (Schaap et al., 2014). Estas herramientas hacen posible diseñar cuestiones que no solamente evalúen la comprensión de conceptos sino la capacidad de los estudiantes para proceder a su aplicación (Ebel y Frisbie, 1979) y son capaces de desplazar a los estudiantes desde una interface de actuación pasiva a una activa (Easdown, 2006).

En el marco de la evaluación continuada de la asignatura Análisis Clínicos y Diagnóstico de Laboratorio del grado de Farmacia de la Universidad de Barcelona se emplea un test objetivo de 50 preguntas con respuesta cerrada verdadero/falso. Se construye una versión inicial del test de la que se derivan dos permutas (A y B) que difieren en el orden de colocación de las preguntas, siendo éstas las mismas en ambas.

El objetivo del presente trabajo es determinar el índice de discriminación de cada *ítem* y si existe o no una influencia de la permuta sobre la calificación obtenida por los estudiantes. Como objetivo adicional se propone determinar la posible influencia de otros factores extrínsecos al instrumento como grupo y/o sexo.

METODOLOGÍA

Las permutas A y B se aplican aleatoriamente a los cinco grupos de estudiantes (M1=82, M2=76, M3=85, T1=84 y T2=31 estudiantes; M=grupos de docencia matinal; T=grupo de docencia de tarde) que conforman la asignatura Análisis Clínicos y Diagnóstico de Laboratorio impartida en el segundo curso, cuarto semestre. Cada permuta es aplicada a 179 estudiantes. Las respuestas son procesadas mediante lector óptico y la calificación de los estudiantes se establece sobre 10 puntos.

Atendiendo a que las preguntas del test son proporcionadas por los diferentes profesores de las diversas áreas que constituyen el contenido de la asignatura, cada uno de ellos se muestra interesado en conocer la capacidad discriminante de las preguntas de que es responsable, para lo cual se determina el índice de discriminación de cada *ítem* en cada una de las permutas utilizadas. El índice de discriminación se calcula mediante la separación de las calificaciones obtenidas por el conjunto de los estudiantes que han respondido cada pregunta en dos grupos: grupo superior que incluye el 27 % de las mejores calificaciones y grupo inferior, incluyendo el 27 % de las peores calificaciones obtenidas (Kelley, 1939; Ros y Weitzmann, 1964). El cálculo implica sustraer el número de aciertos obtenido en una pregunta por el grupo con peores calificaciones del número de aciertos obtenido por el grupo con buenas calificaciones y dividir esta diferencia por el tamaño de un grupo. El rango de este índice es -1 a +1. En general, valores superiores a 0.4 son considerados altos e inferiores a 0.2 son considerados bajos (Ebel, 1954).

Se crea un fichero tipo dBase que recoge los siguientes campos: ITEM: número de orden de la pregunta; R_n: siendo “n” un carácter perteneciente al tipo de respuesta; si en este campo aparece una “X” significa que en la plantilla se ha marcado esa respuesta (por ejemplo, si en el

campo R_A hay una “X” significa que en la plantilla la respuesta correcta es la “A” (verdadero), y R_B indica que la respuesta correcta es “B” (falso); R_n_SUP: lo mismo que en el punto anterior, pero indica cuantos estudiantes marcaron la respuesta “n” a esa pregunta; PCT_AC_SUP: porcentaje de aciertos del grupo superior; PCT_AC_INF: porcentaje de aciertos del grupo inferior; discriminación: este parámetro indica cuanto separa una pregunta a los que saben de los que no saben, y varía entre -1 y +1. Se calcula como: Índice de discriminación = (Sup - Inf)/N, valores extremos -1, 1. La mejor pregunta sería aquella en que acertaran todos los que saben y ninguno de los que no saben. Se clasifica las preguntas según su índice de discriminación: 0.35 o más, pregunta excelente; 0.25 a 0.34, pregunta buena; 0.15 a 0.24, pregunta límite; menos de 0.15, pregunta mala. El valor más alto es la mejor pregunta.

Se compara las clasificaciones obtenidas por las preguntas en las permutas A y B mediante métodos estadísticos adecuados para comparar este tipo de ítems cualitativos (Monleón y Rodríguez, 2017), t-test para comparar las medias entre grupos y el test de Levene para comparar sus varianzas. Adicionalmente, atendiendo a la naturaleza categórica de las variables, se construye una tabulación cruzada o tabla de contingencia que permite el estudio de la relación existente entre dos variables categóricas, en este caso la clasificación (mala, buena, excelente, óptima) obtenida por las preguntas en cada una de las dos permutas A y B.

La posible influencia que pueden ejercer los factores inherentes al instrumento de valoración, como la permuta empleada, o los factores externos a este instrumento como grupo y sexo, o sus interacciones (permuta y sexo; permuta y grupo; grupo y sexo; permuta, grupo y sexo) sobre las calificaciones obtenidas por los alumnos se valoran mediante la construcción de un modelo lineal de evaluación mediante análisis de la varianza (ANOVA), tomando la calificación del estudiante como variable dependiente continua o respuesta y los factores (grupo, sexo) como variables independientes categóricas. Se valida el modelo utilizado mediante un gráfico Q-Q de los residuos. Se ha realizado *a posteriori* un análisis de medias “post-hoc” con el método de Tukey para determinar si existen diferencias entre las calificaciones

$$Y_{ifkl} = \mu + \underset{\text{Grup}}{\alpha i} + \underset{\text{Permuta}}{\beta f} + \underset{\text{Sexe}}{\gamma k} + (\alpha\beta) + (\alpha\gamma) + (\beta\gamma) + (\alpha\beta\gamma) + \text{error}$$

medias de cada grupo por parejas y que tiene en cuenta el denominado “Inflamiento del error tipo I”. Utilizando el resultado “post-hoc” entre parejas de grupos, se ha construido una serie de grupos homogéneos, es decir, una clasificación de grupos que no presentan diferencias entre sí ($p>0.05$) dentro del grupo) (Monleón Getino, 2016).

Todos los cálculos han sido realizados con el paquete estadístico SPSS IBM Versión 23 y las significaciones estadísticas se han establecido con un nivel de $\alpha=0.05$.

RESULTADOS Y DISCUSIÓN

Los resultados obtenidos en el análisis de las dos permutas de acuerdo con los parámetros indicados permiten observar que la misma pregunta puede tener un índice de dificultad y un índice de discriminación diferente en cada una de ellas. Así, la misma pregunta puede ser clasificada como mala en una permuta y como buena en la otra permuta. Consecuentemente, una pregunta específica no puede ser calificada de manera absoluta como buena o mala atendiendo solamente a un cálculo de su índice de discriminación obtenido en un conjunto de preguntas. La capacidad discriminatoria de cada pregunta depende del conjunto de preguntas que integran el instrumento de evaluación.

No obstante, se observa ausencia de diferencias significativas ($p>0,05$) en los índices de dificultad y de discriminación de las preguntas entre las permutas A y B, tomadas en sus conjunto, tras aplicación de t-test ($p=0,666$, $p=0,994$) y Levene ($p>0.05$).

El porcentaje de preguntas malas, buenas, excelentes y óptimas se muestra diferente en la permuta A (54, 18, 24 y 4%, respectivamente) y en la permuta B (46, 20, 34 y 0%, respectivamente). También difiere, en ambas permutas, el porcentaje en que cada una de estas categorías de preguntas (malas, buenas, excelentes y óptimas) se presenta respecto al total de preguntas de su categoría (54, 47,4, 41,4, 100% en la permuta A, respectivamente; 46, 52,6, 58,6 y 0% en la permuta B, respectivamente). La aplicación de la prueba de X^2-cuadrado proporciona un valor $p=0,357$ y muestra, así, la inexistencia de diferencias significativas entre estas distribuciones de preguntas, confirmando la no relación entre las permutas y las clasificaciones de

las preguntas.

Estos resultados permiten indicar que ambas permutas constituyen, cada una de ellas en su conjunto, instrumentos con capacidad evaluatoria carente de diferencias significativas. *A priori,* su aplicación en los distintos grupos y estudiantes de la asignatura ha de ofrecer resultados evaluatorios plenamente comparables y admisibles. Asimismo, sugieren que si bien la misma pregunta puede tener un nivel de dificultad y un índice de discriminación específicos, su inclusión en un conjunto de preguntas puede conducir a resultados de calificación no diferenciables significativamente.

Ahora bien, la frecuencia de las calificaciones obtenidas por los alumnos (número de suspensos, aprobados, notables y excelentes), así como la calificación media correspondiente a cada grupo resulta distinta en los cinco grupos de la asignatura, como consecuencia de diferencias en el número de preguntas acertadas, erróneas y en blanco observadas en los grupos evaluados. El test de Shapiro-Wilk determina la normalidad de la distribución de frecuencias de las calificaciones de los estudiantes en todos grupos ($p<0,05$) a excepción del grupo M3 ($p=0,019$).

El análisis de la significatividad de las diferencias en la calificación obtenida por el estudiante dependiendo de la permuta que haya recibido, del grupo al que pertenece o del sexo, o de las combinaciones entre estos factores, mediante el modelo lineal de evaluación de la varianza (ANOVA) propuesto, indica que únicamente el factor grupo parece ejercer una influencia significativa sobre las calificaciones obtenidas por los estudiantes ($p=0,013$). El resultado prueba que la calificación final del estudiante solamente está influenciada por el grupo y no aparece relacionada con los restantes factores ni con sus combinaciones. Adicionalmente, sólo la combinación de los factores permuta y grupo presenta un $p=0,062$, algo próximo al nivel de significatividad ($\alpha=0.05$). Respecto a la validez del modelo utilizado, se han realizado un gráfico Q-Q de los residuos, que no evidencia desviaciones de la normalidad, si bien el test de Kolmogorov para los residuos presenta un valor de $p=0,020$. Se ha realizado un test de Levene para comprobar la homocedasticidad de las varianzas entre los diferentes niveles de los factores estudiados (ej: sexo, grupo, permuta) y el resultado muestra ($p=0,028$) que no existe una completa

homocedasticidad, si bien la robustez del modelo y la cantidad de datos utilizados nos indican que el p-valor obtenido no habría de variar de p=0,013.

Los resultados parecen indicar que el grupo resulta ser un factor de interés en cuanto a la predicción del éxito académico de los estudiantes. En este sentido cabe destacar que son varios los factores cuya influencia sobre el éxito académico ha sido observada (Mau y Lynn, 2001, Bahar, 2010), como género, soporte social percibido y status sociométrico. La influencia del género puede ser percibida significativamente en aspectos como la implicación o compromiso y el éxito académico (Casuso-Holgado et al., 2013). El grupo puede influenciar los resultados académicos, fundamentalmente atendiendo a su dimensión. Gleason (2012) muestra que los grupos medios (30–55 estudiantes) tienen poca o ninguna ventaja sobre los grupos grandes (110–130 estudiantes) en referencia a los resultados obtenidos por los estudiantes. El área en que se ha demostrado un efecto positivo de los grupos pequeños es en el compromiso o implicación de los estudiantes (Pezzella, et al., 2014). En el presente trabajo, uno de los grupos menos numerosos no obtuvo resultados positivos en sus calificaciones, indicando que, en la situación descrita, el tamaño del grupo no parece influir positivamente en el éxito de los estudiantes.

El análisis de medias “post-hoc” con el método de Tukey proporciona el siguiente esquema de grupos homogéneos (p>0,05 dentro del grupo) para la calificación del estudiante: T1T2 p=0,088; T1M2M3 p=0,120; M2M3M1 p=0,089 (Figura 1). Son significativas las diferencias entre los conjuntos de grupos homogéneos (p<0,05 inter-grupos). Resulta evidente que los grupos cuya docencia tiene lugar por la tarde, grupos T1 y T2, conforman una población homogénea en cuanto a calificaciones obtenidas por sus estudiantes y diferenciada significativamente de las restantes agrupaciones. El grupo T2 constituye un elemento no integrable en ninguna de las agrupaciones que incluyen grupos de docencia de mañana. Por el contrario, el grupo T1 resulta asimilable a los grupos M2 y M3.

5.1. FIGURA O IMAGEN 1

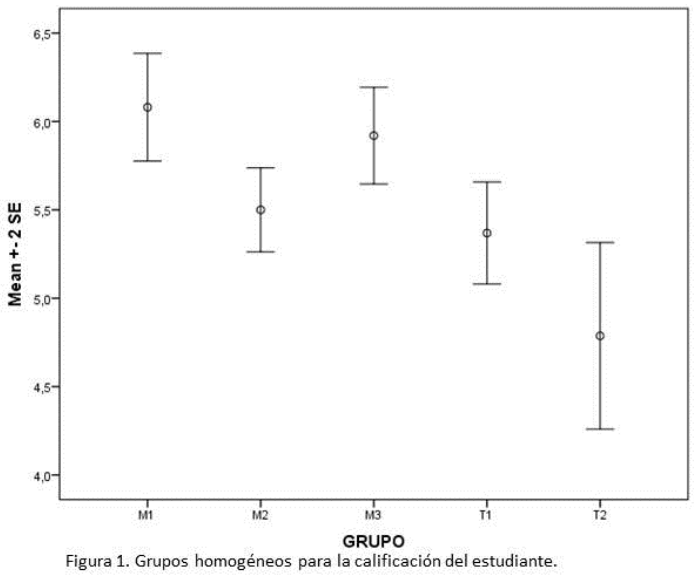

Figura 1. Grupos homogéneos para la calificación del estudiante.

6. REFERENCIAS BIBLIOGRÁFICAS (según normativa APA)

Bahar, H.H. (2010). The effects of gender, perceived social support and sociometric status on academic success. Procedia Social and Behavioral Sciences, 2, 3801–3805

Burton, R.F. (2004). Multiple choice and true/false tests: reliability measures and some implications of negative marking. Assessment and Evaluation in Higher Education, 29, 585-595

Burton, R.F. (2005). Multiple-Choice and True/False Tests: Myths and Misapprehensions. *Assessment and Evaluation in Higher Education*, *30*, 65-72.

Casuso-Holgado, M.J., Cuesta-Vargas, A.I., Noelia Moreno-Morales, N., Labajos-Manzanares, M.T., Barón-López, F.J. y Manuel Vega-Cuesta, M. (2013). The association between academic engagement and achievement in health sciences students. *BMC Medical Education, 13*, 33-39.

Easdown, D. (2006). Integrating assessment and feedback to overcome barriers to learning at the passive/active interface in mathematics courses. *Proceedings of the Science Teaching and Learning Research Symposium*, University of Sydney. UniServe Science, 37-42.

Ebel, R.L. (1954). Procedures for the Analysis of Classroom Tests. *Educational and Psychological* Measurement, 14, 352-364.

Ebel, R.L. y Frisbie, D.A. (2009). *Essentials of educational measurements* (5th edition). New Delhi: Prentice Hall of India.

Gleason, J. (2012). Using technology-assisted instruction and assessment to reduce the effect of class size on student outcomes in undergraduate mathematics courses. *College Teaching, 60*, 87–94

Guba, E. G. y Lincoln, Y. S. (1989). *Fourth Generation Evaluation*. Newbury Park, CA. Sage.

Joint Committee on Standards for Educational Evaluation (1988). *The personnel evaluation standards.* Newbury Park, CA. Sage.

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

Mateo, J. (1993). *La evaluación en el aula universitaria*. Zaragoza: ICE-Universidad de Zaragoza.

Mau, W. C. y Lynn, R. (2001). Gender differences on the scholastic aptitude test the American college test and college grades. *Educational Psychology, 21*, 133-136.

Monleón-Getino, T. (2016). Diseño y planificación de estudios científicos: Calidad de datos (data management) y principios de diseño experimental. Barcelona, Lulú Press Inc.

Monleón-Getino, T. y Rodríguez, C. (2017). Probabilitat i estadística per a ciències II. Barcelona, Edicions i Publicacions UB.

Pezzella, F.S., Paladino, A., Zoller, C. y Mandery, E. (2014). The efficacy of student learning in large-sized criminal justice preparatory classes. *Journal of Criminal Justice Education,* 25, 106-130.

Rodríguez Neira, T., Álvarez Pérez, L., Cadrecha Caparrós, M.A., Hernández García, J., Luengo García, M.A., Ordóñez Álvarez, J.J., Soler Vázquez, E. (1995). *Evaluación de los aprendizajes*. Aula Abierta Monografías 25. Oviedo: ICE-Universidad de Oviedo.

Ross, J. y Weitzman, R.A. (1964). The twenty-seven per cent rule. *The Annals of Mathematical Statistics*, *35*, 214-221.

Schaap, L., Verkoeijen, P. y Schmidt, H. (2014). Effects of different types of true–false questions on memory awareness and long-term retention. *Assessment and Evaluation in Higher Education,*

39, 625-640.

Stufflebeam, D. L. (2001). The metaevaluation imperative. *American Journal of Evaluation, 22*, 183-209.

Tasdemir, M. (2010). A Comparison of Multiple-Choice Tests and True-False Tests Used in Evaluating Student Progress. *Journal of Instructional Psychology, 37*, 258-266.

Tyler, R. W. (1950). *Basic principles of curriculum and instruction*, Chicago, USA: University of Chicago Press.

Tyler, R. W. (1967). Changing concepts of educational evaluation. En R. E. Stack (Comp.), *Perspectives of curriculum evaluation*. AERA Monograph Series Curriculum Evaluation, 1. Chicago, USA: Rand McNally.

Tyler, R. W., (Ed.) (1969). *Educational evaluation: New roles, new means*, Chicago, USA: University of Chicago Press.

Weir, C.J. (2005). Limitations of the Common European Framework for developing comparable examinations and test. *Sage Journals, 22*, 281-300.