Rev Med Chile 2018; 146: 46-52

Análisis de pruebas escritas bajo los principios de la evaluación auténtica. Estudio comparativo entre carreras de la salud y otras carreras de dos universidades de la Región del Biobío

VERÓNICA VILLARROEL[1,a,c], DANIELA BRUNA[1,a], CLAUDIO BUSTOS[2,b], CAROLA BRUNA[3,c,e], CAROLINA MÁRQUEZ[4,d]

[1]Centro de Investigación y Mejoramiento de la Educación (CIME). Facultad de Psicología. Universidad del Desarrollo. Concepción, Chile.
[2]Departamento de Psiquiatría y Salud Mental. Facultad de Medicina. Universidad de Concepción. Concepción, Chile.
[3]Departamento de Bioquímica y Biología Molecular. Facultad de Ciencias Biológicas. Universidad de Concepción. Concepción, Chile.
[4]Departamento de Educación Médica. Facultad de Medicina Universidad de Concepción. Concepción, Chile.
[a]Psicólogo.
[b]Psicólogo, Magíster en Psicología Educacional y Magíster en Estadística Aplicada.
[c]Bioquímico.
[d]Bioquímico, Magíster en Educación Médica para las Ciencias de la Salud.
[d]PhD.

Recibido el 3 de julio de 2017, aceptado el 20 de diciembre de 2017.

Correspondencia a:
Verónica Villarroel
Ainavillo 456. Concepción.
Teléfono: 56412268773.
vvillarroel@udd.cl

Written tests analysis under the principles of authentic assessment. A comparative study of written tests of medical and other undergraduate programs

Background*: Learning assessment has great impact in students´ achievement. However, it is one of the least intervened and researched areas in higher education institutions, all over the world.* ***Aim****: To compare the written tests applied to students of three health science undergraduate programs (Speech Therapy, Medical Technology and Nursing), with the written tests of three programs of other areas (Business and Administration, Psychology and Bioengineering).* ***Material and Methods****: Comparisons were done using the Authentic Assessment Model´s indicators. Also, the magnitude of the change in these variables was evaluated in these two groups of undergraduate programs, after the participation of the teachers in a training program based on this model. A quantitative and repeated measurements design was used. Nineteen teachers participated (nine from medical sciences and 10 from other areas), who drafted 88 written tests before the intervention (which involved 1,318 items) and 93 written tests (that grouped 1,051 items) after it. Items were analyzed using a Hierarchical Lineal Model (HLM), controlling the tests´ and the teachers´ effects.* ***Results:*** *Both groups of undergraduate programs use multiple choice items with a higher frequency, although there were differences in the rest of the items. Also, HLM analysis showed that these programs differed in their changes after the intervention. Health science programs had less improvement in changing the kind of items used, but improved more in Authentic Assessment indicators.* ***Conclusions****: Written tests improved after an intervention aiming to improve the teachers' skills to prepare such tests.*

(Rev Med Chile 2018; 146: 46-52)

Key words: *Educational measurement; Examination questions; Learning; Process assessment (Health Care).*

La evaluación del aprendizaje tiene gran impacto en la calidad de lo aprendido[1,2]. Impulsa y modela los tipos de aprendizajes que se quieren lograr y consolida el uso de ciertas habilidades cognitivas[3], junto con su calidad y profundidad[4-6]. Además, permite a los profesores saber cómo los estudiantes están aprendiendo, apreciar la calidad de su desempeño y tomar decisiones pedagógicas respecto a cómo se seguirá avanzando en el proceso de enseñanza[7].

En la educación chilena, la evaluación del aprendizaje es el área más deficitaria de la práctica pedagógica[8]. En educación superior se ha avanzado, especialmente, en las carreras del área de la salud, a través de tareas de desempeño como portafolios, aprendizaje basado en problemas (ABP), simulaciones o evaluación clínica objetiva estructurada (ECOE)[9,10]. Sin embargo, en general, la apertura de los docentes para hacer cambios en su sistema de evaluación se ha concentrado en la medición de procedimientos o habilidades[11]. Los docentes muestran resistencias para cambiar la evaluación formal a través de pruebas y exámenes escritos, sosteniendo la creencia de que aprender involucra la reproducción literal del saber, la acumulación de datos, conceptos y habilidades básicas. Esta forma de evaluar lo aprendido influencia a los estudiantes en adoptar un enfoque superficial de aprendizaje en comparación a un enfoque profundo[12,13].

Hace más de 20 años, el paradigma de la evaluación avanzó desde los tests objetivos y estandarizados, que se focalizaban en medir porciones de conocimiento atomizado, hacia una evaluación más compleja y amplia de conocimiento y habilidades de orden[14,15]. Desde esta mirada, la evaluación, la enseñanza y el aprendizaje están íntimamente relacionados. Una buena evaluación debe estar diseñada de tal forma que los estudiantes, al rendirla, aprendan y se comprometan con su proceso educativo, permitiéndoles aplicar lo aprendido para resolver algún problema[16].

La evaluación auténtica forma parte de este cambio de paradigma. Busca acercar lo que sucede en las aulas a la vida real, replicando las tareas y estándares de desempeño que típicamente enfrentan los profesionales en el mundo del trabajo[17]. Enfrenta a los estudiantes con problemas que simulan contextos realistas y problematizadores, midiendo habilidades cognitivas de orden superior[18].

La literatura señala, además, que la evaluación auténtica tiene un impacto en la autonomía del estudiante[19], su compromiso y motivación con el proceso de aprender[20], capacidad de autorregulación, metacognición y autorreflexión[21]. Es una metodología pertinente de utilizar en educación superior y que se convierte en una oportunidad para vincular la academia con el mundo del trabajo.

Bajo este modelo, el diseño de una evaluación escrita debe contemplar dos condiciones. Por un lado, medir el aprendizaje de manera contextualizada, a través de estímulos realistas y situados, comprometiendo a los estudiantes con problemas o preguntas importantes, que vale la pena responder más allá del interés pedagógico de las aulas[22]. Por otro lado, debe presentar un nivel de complejidad cognitiva que demande al estudiante a construir conocimiento, haciendo uso de habilidades cognitivas complejas[23]. Se ha comprobado que, cuando los estudiantes hacen uso de este tipo de habilidades, logran mayor profundidad en la comprensión del contenido[4] y estabilidad en el recuerdo de lo aprendido[24].

En este artículo se presenta un estudio en 6 carreras de dos universidades de la Región del Biobío (fonoaudiología, tecnología médica y enfermería, del área de salud e ingeniería comercial, psicología y bioingeniería, de otras áreas). Se realizó un análisis de los ítems de pruebas escritas aplicadas a estudiantes de tercer y cuarto año de carrera, respecto al cumplimiento de los principios de la metodología de evaluación auténtica. Este análisis se hizo antes y después de una intervención, en la que se capacitó a los docentes en la metodología de evaluación auténtica. De esta forma, se compararon los resultados de las carreras de la salud y las de otras áreas, a nivel de pre y posttest. El objetivo del trabajo es describir la realidad de la construcción de pruebas escritas y relevar la necesidad de evaluar los aprendizajes de manera contextualizada, con problemas cercanos a la realidad profesional, que midan habilidades cognitivas de orden superior, dando cuenta de la complejidad cognitiva que requiere la resolución de problemas vinculada al ámbito profesional.

Se intencionó evaluar cómo se construyen las pruebas en carreras de salud, comparando con otro tipo de programas, dado que, generalmente, estas presentan ventajas respecto a la calidad y "autenticidad" de su proceso formativo. Se caracterizan por tener unidades de educación médica,

en las que capacitan a sus profesores en temas de docencia universitaria, y utilizan actividades prácticas en contextos auténticos, como simulaciones y observaciones.

Material y Método

Diseño de investigación

La metodología fue cuantitativa, con un alcance descriptivo y correlacional. El diseño corresponde a un estudio longitudinal de medidas repetidas[25]. El muestreo fue intencionado. El tamaño muestral fue de 2.369 ítems (1.318 en el pretest y 1.051 en el posttest). Estos se encontraban distribuidos en 181 pruebas (88 en el pretest y 93 en el posttest). La carrera que entregó más evaluaciones contaba con 708 ítems y 41 pruebas, y la que entregó menos presentaba 203 ítems y 12 pruebas. El promedio de ítems por carrera fue de 394.83 y el promedio de pruebas por carrera fue de 3.017.

Instrumentos

Se trabajó con una escala de apreciación que medía indicadores de la evaluación auténtica, la cual se construyó posterior a un estudio del estado del arte del constructo. Esta escala fue evaluada por 5 jueces expertos, siendo el acuerdo interjueces de 92%. Se hizo un estudio piloto para analizar la concordancia entre 2 evaluadores al tabular los ítems de 30 pruebas. El coeficiente de Kappa de Cohen obtuvo un valor de 0,82, mostrando alto acuerdo. Se valoró cada ítem en los siguientes 4 aspectos, utilizando una escala que iba de 1 a 3 puntos, en la que 1 implicaba un nivel bajo, 2 uno medio y 3 uno alto:

1. Complejidad cognitiva

Se clasificó cada ítem en el nivel de habilidad cognitiva que se debe hacer uso para dar respuesta a la pregunta. Los niveles son tres: reconocimiento de información (nivel memorístico: 1 y 2 de Bloom[26]), manejo de información (nivel analítico: 3 y 4 de Bloom[26]) y transferencia de información (nivel decisional: 5 y 6 de Bloom[26]).

2. Realismo

Evaluó la contextualización, problematización y realismo presente en la pregunta, es decir, si lo que se preguntaba estaba relacionado con una situación-problema de la vida real o profesional, en que se debía aplicar el contenido curricular para responder.

3. Alineación con competencias específicas del perfil de egreso

Se analizó la relación de lo medido con las competencias específicas (propias de la profesión) declaradas en el perfil de egreso de cada carrera.

4. Alineación con competencias genéricas del perfil de egreso

Se analizó la relación de lo medido en cada ítem de la prueba con las competencias genéricas (transversales a distintas profesiones) declaradas en el perfil de egreso de cada carrera. Se decidió trabajar con estos indicadores debido a que la revisión de la literatura evidenció que son las cuatro dimensiones más representativas del Modelo de Evaluación Auténtica.

Procedimiento

Se invitó a participar a las carreras (3 de cada universidad), todas con una duración de 5 años, solicitándoles las pruebas escritas de los dos últimos años de calendario, en los niveles de 3er y 4to año de cada carrera, de 19 profesores (4 de psicología, 3 de ingeniería comercial, 3 de bioingeniería, 2 de enfermería, 4 de tecnología médica y 3 de fonoaudiología). También se pidió el perfil de egreso de cada programa. Luego, los docentes fueron capacitados durante un semestre (en 2014), en la metodología de evaluación auténtica y aplicaron lo aprendido durante el siguiente semestre académico[27]. Luego de implementada la metodología, los docentes entregaron las pruebas aplicadas a los estudiantes ese semestre post intervención (mediados de 2015). Las pruebas aplicadas fueron sumativas y de proceso. Correspondían a certámenes y exámenes.

Para asegurar la calidad técnica de los ítems y las pruebas, junto con su relación con los objetivos de los cursos, se tomaron algunos resguardos. Los profesores que participaron en la investigación, y fueron capacitados, fueron sugeridos por los directores de carrera de cada universidad, considerando a docentes que se destacaban en su ejercicio profesional, por su compromiso, experiencia y evaluación docente. Además, las pruebas, construidas a partir de la capacitación, fueron visadas por cada carrera (equipos docentes internos de cada carrera, que se encargan de

procesos académicos y curriculares) respecto al cumplimiento con los contenidos del curso y los resultados de aprendizaje del programa de asignatura que cada profesor dictaba. Finalmente, al construir la prueba, el profesor completaba una lista de chequeo que solicitaba alinear los ítems con los contenidos del curso y los resultados de aprendizaje que la asignatura comprometía, como también la redacción de la pregunta.

Plan de análisis de datos

El análisis de datos se efectuó mediante dos procesos. El primero consistió en examinar el tipo de ítems utilizados en las pruebas antes y después de la capacitación, en las carreras de la salud y las carreras de otras áreas. Además, se comparó el cambio en el tipo de ítems, según el área disciplinar de las carreras (salud y otras). Para esto se utilizaron porcentajes.

En el segundo proceso se evaluó el nivel de presencia de los indicadores del Modelo de Evaluación Auténtica en las pruebas antes y después de la capacitación, en las mismas carreras. De igual manera se comparó el cambio según área disciplinar. Para este fin se utilizaron las medias.

Para modelar el cambio en los ítems, en ambos procesos, se empleó como técnica de análisis de datos el Modelo Jerárquico Lineal, HLM[28], anidando los ítems a nivel de la prueba y del profesor.

Aspectos éticos

El Comité de Ética de Investigación Institucional de la Universidad del Desarrollo realizó una revisión del proyecto de investigación, aprobando su ejecución. Posteriormente, se solicitó permiso a las autoridades de todos los programas de pregrado que conformaron la muestra. Finalmente, se invitó a participar a los docentes, quienes fueron informados de la investigación y sus objetivos a través de un consentimiento informado. Su firma implicaba que aceptaban entregar sus pruebas, antes y después de la intervención, para ser analizadas. En este también se explicitaba el carácter voluntario de la participación y se aseguraba el resguardo del anonimato de los datos.

Resultados

Los resultados evidencian que antes de la capacitación, los ítems más utilizados, por ambos tipos de carrera, eran de selección múltiple, seguidos por verdadero y falso, desarrollo breve y análisis de casos, en las carreras de la salud. En las otras carreras (ingeniería comercial, psicología y bioingeniería), los más utilizados fueron análisis de caso, desarrollo breve y desarrollo extenso. Si se agrupa los ítems en preguntas de respuesta abierta y cerrada, las carreras de la salud superaban a las carreras de otras áreas en ítems de respuesta cerrada, mientras que las demás carreras destacaban en su nivel de preguntas de respuesta abierta (Tabla 1).

Posterior a la intervención, en las carreras de la salud se utilizó más análisis de caso, seguido por desarrollo breve, selección múltiple, y verdadero o falso, mientras que en las demás, se usó más

Tabla 1. Porcentaje de tipos de ítems utilizados en las pruebas, según área disciplinar de las carreras

Tipo de ítem	% Carreras de la salud		% Carreras de otras áreas		Valor-p[c]
	Pretest 2014	Posttest 2015	Pretest 2014	Posttest 2015	
Análisis de caso[a]	17,08	33	14,42	20,24	0,001*
Desarrollo breve[a]	18,86	23,10	13,66	33,66	< 0,001*
Desarrollo extenso[a]	0,16	3,31	12,94	10,74	0,5
Resolución breve de problemas[a]	0,49	4,22	5,72	1,47	0,99
Selección múltiple[b]	29,59	19,37	49,56	33,89	0,04*
Completación[b]	6,18	5	0	0	0,02*
Verdadero y falso[b]	27,64	12	3,70	0	< 0,001*

[a]Ítems de respuesta abierta; [b]Ítems de respuesta cerrada; [c]Valor-p alude a la significancia del cambio logrado en ambos grupos de carreras; *Valor-p significativo.

Tabla 2. Indicadores de Evaluación Auténtica medidos en las pruebas, según área disciplinar de la carrera

Tipo de ítem	Carreras de la salud		Carreras de otras áreas		Valor- p[c]
	Antes 2014	Después 2015	Antes 2014	Después 2015	
Complejidad cognitiva	1,41	1,75	1,61	1,86	0,03*
Realismo	1,45	2,04	1,53	1,82	< 0,001*
Alineación con competencias específicas del perfil de egreso	1,39	1,78	1,57	1,89	0,02*
Alienación con competencias genéricas del perfil de egreso	1,27	1,58	1,14	1,30	0,04*

[c]Valor-p alude a la significancia del cambio logrado en ambos grupos de carreras; *Valor-p significativo.

selección múltiple, seguido por desarrollo breve, análisis de caso y desarrollo extenso. También se observó un cambio relevante en los ítems en preguntas de respuesta abierta y cerrada, tras la intervención. Las carreras de la salud disminuyeron a la mitad sus ítems de respuesta cerrada y aumentan al doble los de respuesta abierta, logrando que no existieran diferencias significativas entre ambos tipos de ítems con las otras carreras.

Por otro lado, el análisis realizado con HLM mostró que existían cambios dependientes del tipo de carrera en todos los ítems, a excepción de los de resolución breve de problemas y de desarrollo extenso. En análisis de caso, se observa que las carreras de salud aumentaron en más de 16%, siendo este aumento mucho mayor que el de las demás carreras. Lo mismo ocurrió con la utilización de desarrollo breve. Su uso aumentó de mayor manera en las otras carreras. Si bien, se disminuyó el uso de selección múltiple, esta baja fue mucho mayor en las carreras de otras áreas. Algo similar ocurrió con los verdadero o falso y la completación, los cuales disminuyeron, pero en las demás carreras no se utilizaron más (Tabla 2).

Antes de la capacitación se observaba que todos los indicadores se encontraban cercanos al puntaje medio, a excepción de la alineación con las competencias genéricas del perfil de egreso, que fue el indicador con una media más baja. Posterior a la intervención se evidenció que todos los indicadores aumentaron sus puntajes. Sin embargo, la alineación de los ítems con las competencias genéricas sigue siendo el más bajo, para ambos tipos de carreras.

El análisis con HLM mostró que existieron cambios dependientes del tipo de carrera en todos los indicadores. Todas las carreras aumentaron sus puntajes posterior a la intervención. Sin embargo, el aumento fue mayor en las carreras de la salud, en los indicadores de complejidad cognitiva, realismo y alineación con las competencias específicas, mientras que el aumento fue mayor para las carreras de otras áreas en el indicador de alineación con las competencias genéricas.

Discusión

Las carreras de la salud tienen gran experiencia a la hora de evaluar competencias, habilidades o procedimientos, a través de tareas de desempeño como son las simulaciones, ECOE u observaciones. No obstante, la construcción de pruebas escritas, que representan la forma de evaluar el contenido curricular, presenta importantes oportunidades de mejora, al compararla con carreras de otras áreas disciplinares. Antes de la intervención, las pruebas presentan un alto número de ítems de respuesta cerrada, descontextualizados y de baja complejidad cognitiva. Esta forma de preguntar acerca del conocimiento, tiende a favorecer un aprendizaje superficial y memorístico en los estudiantes, dificultando la aplicación del saber. Las pruebas escritas carecen de autenticidad, estando desvinculadas con los problemas que se enfrentan en el mundo del trabajo. A pesar de ello, la fortaleza de las carreras de la salud es que sus profesores están abiertos a

aprender rápidamente y son capaces de generar cambios en la construcción de pruebas escritas, al ser capacitados para ello. Una vez realizado el entrenamiento docente en la metodología de evaluación auténtica, las pruebas escritas cambiaron significativamente, aumentando el número de ítems de respuesta abierta, disminuyendo las preguntas de respuesta cerrada, aumentando la complejidad cognitiva y realismo de los ítems, como también su nivel de alineación con las competencias específicas del perfil de egreso de cada carrera. De esta forma, lograron equipararse a los índices de las carreras de otras áreas, e incluso superarlos, por ejemplo, en el realismo y contextualización de los ítems.

Referencias

1. Boud D. Sustainable assessment: rethinking assessment for the learning society. Studies in Continuing Education 2010; 22 (2): 151-67.
2. Edström K. Doing course evaluation as if learning matters most. Higher Education Research & Development 2008; 27 (2): 95-106.
3. Vu T, Dall´Alba G. Authentic Assessment for Student Learning: An ontological conceptualization, Educational Philosophy and Theory 2014; 46 (7): 778-91.
4. Jensen J, McDaniel M, Woodard S, Kummer T. Teaching to the test… or testing to teach: exams requiring higher order thinking skills encourage greater conceptual understanding. Educ. Psychol. Rev 2014; 26: 306-29.
5. Syafei M. Backwash effects of portfolio assessment in academic writing classes. TEFLIN Journal 2012; 23 (2): 206-21.
6. Watkins D, Dahlin B, Ekholm M. Awareness of backwash effect of assessment: A phenomenographic study of the views of Hong Kong and Swedish lectures. Instructional Science 2005; 33: 283-309.
7. Wiliam D. Keeping learning on track: Formative assessment and the regulation of learning. In *Second Handbook of mathematics teaching and learning*. F.K. Lester Jr., Editors. Greenwich, CT: Information Age 2007. p. 1053-98.
8. Manzi J, González R, Sun Y. (Eds.) *La evaluación docente en Chile*. Santiago, Chile: Facultad de Ciencias Sociales, Escuela de Psicología, PUC. 2011.
9. O´haja M, Dunlea M, Muldoon K. Group marking and peer assessment during a group poster presentation: the experiences and views of mindwifery students. Nurse Education in Practice 2013; 13 (5): 466-70.
10. Wu X, Heng M, Wang W. Nursing students´ experiences with the use of authentic assessment rubric and case approach in the clinical laboratories, Nurse Education Today 2015; 35: 549-55.
11. Biggs J, Tang C. *Teaching for quality learning at university: What the student does*. Maidenhead, Berkshire, Open University Press. 2011.
12. Beyaztas DI, Senemoglu N. Learning approaches of successful students and factors affecting their learning approaches. Education and Science 2015; 40 (179): 193-216.
13. Endedijk MD, Vermunt JD. Relations between student teachers' learning patterns and their concrete learning activities. Studies in educational evaluation 2013; 39 (1): 56-65.
14. Birenbaum M. New insights into learning and teaching and their implications for assessment. In: Segers M, Dochy F, Cascallar E. Editors, *Optimizing New Modes of Assessment: In Search of Qualities and Standards*. Dordrecht: Kluwer Academic Publishers; 2003. p. 13-36.
15. Wiliam D. What is assessment for learning? Studies in Educational Evaluation 2011; 37: 3-14.
16. Gulikers J, Kester L, Kirschner P, Bastiaens T. The effect of practical experience on perceptions of assessment authenticity, study approach, and learning outcomes. Learning and Instruction 2008; 18: 172-86.
17. Wiggins G. The case for authentic assessment. Practical Assessment, Research & Evaluation 2011; 2 (2).
18. Frey B, Schmitt V, Allen J. Defining authentic classroom assessment. Practical assessment, Research & Evaluation 2012; 17 (2): 1-18.
19. Raymond J, Homer C, Smith R, Gray J. Learning through authentic assessment: An evaluation of a new development in the undergraduate midwifery curriculum. Nurse Education in Practice 2012; 13: 471-6.
20. Nicol D, Thomson A, Breslin C. Rethinking feedback practices in higher education: a peer review perspective, Assessment & Evaluation in Higher Education 2014; 39 (1): 102-22.
21. Ling Lau K. Chinese language teachers' perception and implementation of self-regulated leaning-based instruction. Teacher and Teaching Education 2013; 31: 56-66.
22. Wiggins G, McTighe J. Examining the Teaching Life. Educational Leadership 2006.
23. Thornburn M. Articulating a Merleau-Pontain phenomenology of physical education: The quest for active student engagement and authentic assessment in high-stakes examination awards. European Physical Education Review 2008; 4 (2): 263-80.

24. Rawson K, Dunlosky J, Sciartellli S. The power of successive relearning: improving performance on course exams and long term retention, Educational Psychology Review 2013; 25: 523-48.
25. Andreb H, Golsch K, Schmidt A. *Applied Panel Data Analysis for Economic and Social Surveys*. Editor Springer Science & Business Media. Cologne, Germany. 2013.
26. Bloom B, Masia B, Krathwohl D. *Taxonomy of Educational Objectives*. New York: McKay. 198.
27. Villarroel V, Bruna D, Bruna C, Herrera C, Márquez C. Implementación de la Metodología de Evaluación Auténtica en Educación Superior. En Jerez, O, Editor, *Innovando en Educación Superior: Experiencias Claves en Latinoamericana y Caribe*. Volumen 2: Metodologías Activas de Enseñanza y Aprendizaje; 2017.
28. Raudenbuch S, Bryk A. *Hierarchical Linear Models. Applications and Data Analysis Methods*. California: Sage Publications. 2002.