Psicología Educativa 20 (2014) 61-63

Psicología Educativa

www.elsevier.es/psed

Bringing added value to educational assessment: A shift from an audit mode of assessment to an assistence mode

Cómo aportar valor añadido a la evaluación: de la auditoría a una función asistencial en la educación

María J. Navas*

Universidad Nacional de Educación a Distancia, Spain

Currently, schools are being asked to participate in a higher number of educational assessment programs, with minimal returns, apart from the snapshot of performance about the educational outcomes of a state or country and how it compares to other states or countries. Somehow we are over-tested but under-assessed: «Many schools continue to engage in summative testing-educational autopsies that seek to explain how the patient died but offer no insight to help the patient improve» (Reeves, 2006, p. ix). The scores on these assessments help inform educational policy makers but are of limited value in practical instructional settings, since they do not usually help inform classroom instruction and learning. Is there a way to increase the impact of assessment on learning? Is it possible to develop instructionally-relevant assessment measures?

Randy Bennett has been leading an innovative research program launched by the Educational Testing Service in 2007, whose goal is not only the assessment *of* learning but also *for* learning and *as* learning: CBAL (Cognitively Based Assessment of/for/as Learning). CBAL combines the summative (*of* learning) and formative (*for* learning) components of assessment, working with innovative tasks viewed by teachers and students as worthwhile learning experiences in and of themselves (*as* learning), delivered primarily by computer. CBAL is also a good example of a theory-driven program: the CBAL assessments are developed using a rich theoretical framework with competency models that delineate the knowledge, processes, strategies, and habits of mind important for success in a domain, as well as how students progress from simple to complex performances in specific skills, formulating hypothesized learning progressions (Bennett, 2010).

Designing test items according to a cognitive model has been recognized as an important way to improve the quality of test items and the validity of inferences drawn from test scores (Embretson & Gorin, 2001; Mislevy, 2006; Nichols, 1994). This is because a cognitive model provides an explicit understanding of the knowledge and skills normally used by students to solve standardized tasks in a test domain. Unfortunately, there is still a shortage of models that characterize student performance in most testing situations. As part of the CBAL English Language Arts competency model, Paul Deane and Yi Song (this issue) present a case study where they describe the framework and learning progressions for argumentation, a complex skill playing an important role not only in reading and writing but in everyday life. Peter van Rijn, Aurora Graf, and Paul Deane later address the empirical recovery of the learning progressions of this skill for middle school students, scrutinizing their performance on three parallel scenario-based assessment forms, and providing a method that consistently classifies students in the levels of the argumentation learning progression when different forms are used.

Since cognitive models help clarify the psychology that underlies test performance, scores from cognitively-based tests may be more interpretable and meaningful. The remaining papers in this issue describe three approaches that contribute to building the interpretation/use argument and the validity argument (Kane, 2013): the assessment triangle, the Evidence-Centered Design (ECD), and the Cognitively Diagnostic Assessment (CDA) and the Cognitive Diagnosis Model (CDM) framework.

The assessment triangle supplies a framework to improve assessment through its three foundational elements (cognition, observation, interpretation), working with a design process that connects the three elements of the triangle to ensure that the theory of cognition and learning, the observations and the interpretation process work together to support the intended inferences from test scores. The assessment triangle was proposed by the National Research Council (NRC) committee in charge of reviewing the advances in the cognitive and measurement sciences at the beginning of the new millennium (Pellegrino, Chudowsky, & Glaser, 2001). James Pellegrino (this issue) outlines the main ideas underlying this framework, and also considers the use of an ECD process to develop and interpret assessments, as well as the assessment framework driven by models of learning expressed as learning progressions. He advocates a coherent assessment system, describing the contexts and purposes of educational assessment, and the types of assessments required to fulfill the learning goals involved in the process of educational transformation in the new century. He considers assessment as a positive influence on teaching and learning, as long as it is properly conceived, designed, and implemented.

ECD is a way of formalizing the test design process strongly related to the validity of the test scores. It is based on the principles of evidentiary reasoning (Mislevy, Steinberg, & Almond, 2003), and

*Correspondence concerning this article should be addressed to Maria J. Navas. Departamento de Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universidad Nacional de Educación a Distancia. Juan del Rosal, 10. 28040 Madrid (España). E-mail: mjnavas@psi.uned.es

DOI: http://dx.doi.org/10.1016/j.pse.2014.11.005

1135-755X/ © 2014 Colegio Oficial de Psicólogos de Madrid. Producido por Elsevier España, S.L. Todos los derechos reservados.

focuses on what we want to say about test takers (claims) and what evidence we need to allow us to say. Interestingly, the design architecture further ensures coordination of the work from different specialists, such as statisticians, task authors, delivery-process developers, and interface designers. Michael J. Zieky provides an overview of this approach in a paper full of practical and useful recommendations that are the fruits of his long experience in the field.

Jimmy de la Torre and Nathan Minchen summarize the Cognitively Diagnostic Assessment, that uses a cognitive model to guide test design and analysis and provides useful diagnostic information about students' strengths and weaknesses for tailored instruction or remediation purposes, working with Cognitive Diagnostic Models (CDM). They use the triangle assessment and ECD frameworks to introduce the components of such an assessment, and they also detail some CDMs.

As Anastasi (1986) rightly stated, validation efforts must begin at the very beginning of the test design process. However, despite being the most important psychometric property of test scores, validity has not often received as much attention as it deserves (Brennan, 2001; Ebel, 1961). According to Kingston (2007), the future challenges to psychometrics are validity, validity, and validity, that are «the three most important factors contributing to the value of a testing program» (p. 1111). This is also the main point raised by Michael Kane and Isaac Bejar, who were in charge of closing this special issue and wrote the epilogue to this volume: no matter how principled an approach to test design may be, validation is still required, and claims for instructional effectiveness need to be evaluated.

All the former approaches contribute to put a higher value on validity at the outset of the process, and follow Messick's dictates (1994): one must design an assessment from the very start around the inferences one wants to make, the observations one needs to ground them, the situations that will evoke those observations, and the chain of reasoning that connects them. In short, it is all about asking clear questions and providing cogent answers that are supported by logic and evidence (Brennan's Socratic validation) or, otherwise stated, about making claims and evaluating the credibility/plausibility of these claims, and the underlying assumptions (Kane's argument-based approach). Relying more heavily on theories of cognition and learning can help to build the interpretation/use argument and contribute decidedly to increasing the impact of assessment on learning.

«Accountability must be achieved in a way that supports high quality teaching and learning», Dr. Gordon said when presenting the recommendations made by the Gordon Commission (2013) after several years working on what would be needed from educational measurement during the 21st century. «Our conviction is that while the field of measurement in education has established a splendid history primarily directed at the measurement of education, the future of assessment in education will depend on the field's capacity to pursue assessment for education». To some extent this is also the conclusion reached by the NRC Committee on Incentives and Test-Based Accountability in Public Education (National Research Council, 2011). High-stakes testing is not an effective lever for improving teaching and learning: only small, modest effects – and in many cases no effect at all – have been found on student learning, after having carefully studied 15 programs both inside and outside U.S. including the large-scale policies of No Child Left Behind, and state high school exit exams. A shift from an audit mode of assessment to an assistance mode could be extremely helpful to increase teaching/teachers effectiveness but also to bring added value to assessment.

To do so, «what is needed is more attention to the confluence of psychometrics, cognitive psychology, development psychology, learning theory, curriculum, and other theoretical bases of our assessment blueprints. It is not enough to develop parsimonious models – we need models that support appropriate actions… that not only provide accurate diagnoses, but also provide valid prescription leading to demonstrable educational improvement» (Kingston, 2007, p. 1111). CBAL has shown that it is possible to better align assessment with classroom instruction, by incorporating tasks that model teaching and learning practices, tasks that inspire work worth doing by teachers and students. Leighton (2013) claims that «even in the absence of a cognitive model, important steps can be incorporated in the test design process to accommodate cognitive-psychological principles» (p. 20). She shows three examples illustrating a basic change in test philosophy and purpose (SAT, Scholastic Aptitude Test), in the type of validity evidence secured to support test item design (PISA, Programme for International Student Assessment), and in test specifications (BEAR, Berkeley Evaluation and Assessment Research).

Resumen

Actualmente se pide a los centros escolares que participen en un número cada vez mayor de evaluaciones educativas y no es mucho lo que reciben a cambio, aparte de la foto fija que permite comparar los resultados de una región o país con los de otras comunidades o naciones. Según Reeves (2006), "muchas escuelas continúan embarcándose en evaluaciones sumativas, que son autopsias educativas que tratan de explicar de qué murió el paciente pero que no sirven para ayudar a que el paciente mejore" (p. ix). Los resultados en este tipo de evaluaciones ofrecen información de interés para las autoridades educativas pero resultan de escasa utilidad en el contexto escolar aplicado, ya que habitualmente no sirven para retroalimentar el proceso de enseñanza-aprendizaje a pie de obra en el aula. ¿Existe alguna forma de incrementar el impacto de la evaluación en el aprendizaje? ¿Es posible diseñar pruebas con preguntas o ítems que sean sensibles o relevantes a la instrucción?

Los tres primeros trabajos de este número ofrecen una introducción a tres marcos conceptuales que proporcionan alguna respuesta a los interrogantes anteriores: el triángulo de la evaluación, el diseño centrado en la evidencia y el marco de la evaluación para el diagnóstico cognitivo.

El triángulo de la evaluación propone un proceso de trabajo que conecta los tres vértices del triángulo (cognición, observación e interpretación) para garantizar que las teorías cognitivas y del aprendizaje, las observaciones recogidas y la posterior interpretación de las puntuaciones asignadas a las mismas operan sinérgicamente para poder realizar las inferencias deseadas a partir de esas puntuaciones. Este marco fue propuesto por el comité del National Research Council encargado de revisar al comienzo del nuevo milenio los avances acaecidos en el campo de las ciencias cognitivas y de la medición (Pellegrino, Chudowsky y Glaser, 2001). En este número James Pellegrino esboza las principales ideas del triángulo de la evaluación y realiza también algunas consideraciones acerca del diseño centrado en la evidencia, así como del marco de evaluación basado en modelos de progresiones de aprendizaje. Defiende la necesidad de trabajar con un sistema de evaluaciones coherente, describiendo los distintos contextos y objetivos de la evaluación educativa y los distintos tipos de evaluación que se necesitan para dar respuesta a los objetivos de aprendizaje planteados en el proceso de transformación educativa del nuevo siglo. Considera que la evaluación puede influir positivamente en el proceso de enseñanza-aprendizaje siempre que ésta sea adecuadamente concebida, diseñada e implementada.

El diseño centrado en la evidencia proporciona una vía para formalizar el proceso de diseño de un test que está estrechamente relacionada con la validez de sus puntuaciones y que descansa en los principios del razonamiento basado en la evidencia: se centra básicamente en qué es lo que queremos decir acerca de las personas que han respondido al test y qué tipo de evidencia necesitamos para ello. Michael J. Zieky proporciona una revisión de este tipo de diseño en un artículo que contiene numerosas recomendaciones muy útiles y prácticas, fruto de su dilatada experiencia en el campo.

Jimmy de la Torre y Nathan Minchen revisan en este número el marco de la evaluación para el diagnóstico cognitivo, que utiliza un modelo cognitivo para guiar el diseño del test y el posterior análisis de sus puntuaciones y proporciona información diagnóstica muy útil para detectar los puntos fuertes y débiles de los estudiantes, que permitirá tomar las correspondientes decisiones para adaptar la instrucción a cada estudiante individual o a la composición del aula. En este marco se propone trabajar con modelos de diagnóstico cognitivo en lugar de las habituales teorías de tests, presentando los detalles de algunos de estos modelos. Los autores utilizan el triángulo de la evaluación y el diseño centrado en la evidencia para presentar los principales componentes de una evaluación de este tipo.

Los dos siguientes trabajos de este número están firmados por miembros del equipo CBAL (acrónimo inglés de 'evaluación cognitiva *de*, *por* y *para* el aprendizaje'). Se trata de un proyecto que, desde una sólida base teórica, proporciona un excelente ejemplo de cómo se puede combinar el componente formativo y sumativo de la evaluación, al trabajar con ítems que proporcionan tareas que constituyen en sí mismas valiosas experiencias de aprendizaje y que, de algún modo, pueden servir para sugerir o modelar buenas prácticas en el aula.

Utilizando el diseño centrado en la evidencia, Paul Deane y Yi Song presentan una evaluación basada en escenarios en la que se mide la capacidad de argumentar, que juega un papel decisivo no solo en la competencia lecto-escritora sino en la vida diaria. Los autores trabajan con un marco donde se combinan las fases definidas para la argumentación con las progresiones de aprendizaje formuladas para dicha habilidad. En el siguiente trabajo Peter van Rijn, Aurora Graf y Paul Deane replican empíricamente los niveles de estas progresiones de aprendizaje de la argumentación en una muestra de estudiantes de enseñanza secundaria obligatoria; proponen también un método que permite clasificar de manera consistente a los estudiantes en estos niveles trabajando con formas paralelas de la prueba.

En suma, estos dos trabajos ponen de manifiesto que es posible diseñar preguntas que sean sensibles a la instrucción. Los tres trabajos anteriores apuntan a estrategias que permiten abordar el diseño del test y el posterior análisis de sus puntuaciones de forma que se contribuya a construir el argumento de interpretación/uso del test y, subsiguientemente, a forjar su argumento de validez, ya que ponen el acento en la validez desde el inicio mismo del diseño de la prueba. Según Kingston (2007), los retos a los que se enfrenta la psicometría en el futuro son la validez, la validez y, de nuevo, la validez, que son 'los tres factores más importantes que contribuyen al valor de un programa de evaluación' (p. 1111). En esta misma dirección discurre la idea central del epílogo escrito para este número por Michael Kane e Isaac Bejar: por más que se utilice un diseño de evaluación convenientemente anclado en un modelo cognitivo o en una progresión de aprendizaje, es imprescindible validar el uso o interpretación de los resultados de dicha evaluación, esto es, es preciso recabar la evidencia necesaria que permita utilizar o interpretar del modo previsto dichos resultados, así como la evidencia que permita concluir su eficacia en la instrucción.

Después de varios años de trabajo para ver qué se necesita para afrontar los desafíos de la educación del siglo XXI, la comisión Gordon (2013) señalaba que si bien "el campo de la medición educativa ha realizado un trabajo espléndido dirigido fundamentalmente a la medida de la educación, el futuro de la evaluación educativa pasa por la capacidad de este campo para ocuparse de la evaluación para la educación". Este cambio desde una concepción de auditoría de la evaluación a un modo asistencial puede proporcionar ese valor añadido tan necesario para la evaluación, ya que puede contribuir a mejorar la eficacia del proceso de enseñanza-aprendizaje en el aula, al alinear la evaluación con la instrucción trabajando con diseños que, como el proyecto CBAL, integran modelos cognitivos o de aprendizaje en el diseño de la prueba que realizan, además, de una manera muy fundamentada.

Acknowledgements

I wish to thank the authors for their willingness to join this project and for their contributions. I would also like to acknowledge the helpful reviews of Alan Schoenfeld, Russell Almond, Lou DiBello, Deanna Kuhn, and Diana Wilmot. Jose Antonio León, the editor of the journal, trusted me and the project from the very beginning, and the Elsevier staff (Marisa del Barrio, Jose Maria Puente, and Francisco Javier de Sande) made this volumen possible in a careful and fast way.

References

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*, 70-91.

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues & Practice, 20*(4), 6-17.

Deane, P., & Song, Y. (2014). A Case Study in Principled Assessment Design: Designing assessments to measure and support the development of argumentative reading and writing skills. *Psicología Educativa, 20*, 99-108.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*, 89-97.

Ebel, R. (1961). Must all tests be valid? *American Psychologist, 16*, 640-647.

Embretson, S., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*, 343-368.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1-73.

Kane, M., & Bejar, I. (2014). Cognitive frameworks for assessment, teaching, and learning: A validity perspective. *Psicología Educativa, 20*, 117-122.

Kingston, N. (2007). Future challenges to psychometrics: Validity, validity, validity. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics*. Boston, MA: North Holland.

Leighton, J. (2013). Large-scale assessment design and development for the measurement of student cognition. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large-scale assessment in education: Theory, issues, and practice* (pp. 13-26). New York: Routledge.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13-23.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. Brennan (Ed.), *Educational Measurement*. Westport, CT: Praeger and American Council on Education.

Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3-62.

National Research Council. (2011). *Incentives and Test-Based Accountability in Education. Committee on Incentives and Test Based Accountability in Public Education*. M. Hout and S.W. Elliott, Editors. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessment. *Review of Educational Research, 64*, 575-603.

Pellegrino, J. (2014). Assessment as a Positive Influence on 21st Century Teaching and Learning: A systems approach to progress. *Psicología Educativa, 20*, 65-77.

Pellegrino, J. W, Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Reeves, D. B. (2006). Foreword. In L. B. Ainsworth & D. J. Viegut, *Common formative assessment*. Thousand Oaks, CA: Corwin Press.

The Gordon Commission on the Future of Assessment in Education (2013). *Report: Future K–12 Education Assessments Must Help Improve Teaching and Learning, Inform Accountability*. Press release, March, 11.

Van Rijn, P. W., Graf, E. A., & Deane, P. (2014). Empirical recovery of argumentation learning progressions in scenario-based assessments of english language arts. *Psicología Educativa, 20*, 109-115.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa, 20*, 79-87.