PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2015, 52(2), 34-54

Pontificia Universidad Católica de Chile
Facultad de Educación

Evaluación válida de la escritura desde la perspectiva de las comunidades de investigación en escritura y medición

Valid Writing Assessment from the Perspectives of the Writing and Measurement Communities

[1]Nadia Behizadeh y [2]George Engelhard, Jr.

[1]College of Education, Georgia State University, EE.UU.
[2]Department of Educational Psychology, The University of Georgia, EE.UU.

Resumen

Este estudio examina el concepto de validez en dos comunidades de práctica distintas: la de investigación en escritura y la de medición educacional. Las conceptualizaciones de validez han evolucionado diferencialmente dentro de cada una de ellas. Tres preguntas guían nuestro estudio: (a) ¿En qué consiste una evaluación válida de escritura según la comunidad de investigación en escritura? (b) ¿En qué consiste una evaluación válida de escritura según la comunidad de investigación en medición? (c) ¿Cuáles son los puntos de consenso y desacuerdo sobre el concepto de validez en ambas comunidades? El presente estudio busca fomentar la comunicación entre estas dos comunidades académicas con respecto a los problemas asociados con la validez en la evaluación de la escritura. También destacamos las contribuciones de la teoría de medición Rasch (Rasch, 1960/1980) a la comprensión y evaluación de la validez. Nuestras metas son fortalecer la conceptualización de la validez en la evaluación de la escritura e identificar áreas de consenso y disenso en las definiciones de validez existentes. Estos análisis expanden el trabajo previo de Engelhard y Behizadeh (2012), el cual exploró definiciones consensuadas de validez. El presente estudio tiene implicaciones para mejorar la investigación, la teoría y la práctica en la evaluación de la escritura.

Palabras clave: validez, validez consecuencial, evaluación de la escritura, comunidades de práctica, teoría de medición Rasch

Correspondencia a:
Nadia Behizadeh
Department of Middle and Secondary Education, College of Education
Georgia State University
P.O. Box 3978, Atlanta, GA 30302, USA.
Correo electrónico: nbehizadeh@gsu.com
Una versión anterior de este manuscrito fue presentada en el International Objective Measurement Workshop [Taller Internacional de Medición Objetiva] en Filadelfia, PA, abril de 2014.

ISSN: 0719-0409 DDI: 203.262, Santiago, Chile
doi: 10.7764/PEL.52.2.2015.3

Abstract

This study examines the concept of validity in two distinct communities of practice: the writing research and educational measurement communities. Conceptualizations of validity have evolved differentially within each community. Three questions guide our research: (a) What is a valid writing assessment from the perspective of the writing community? (b) What is a valid writing assessment from the perspective of the measurement community? (c) What are some points of consensus and disagreement over the concept of validity in the two communities? This study aims to foster communication between these two different scholarly communities regarding validity issues in writing assessment. We also highlight the contributions that Rasch measurement theory (Rasch, 1960/1980) brings to understanding and evaluating validity. Our goals are to enhance the conceptualization of validity in writing assessment and to identify areas of consensus and disagreement regarding definitions of validity. These analyses extend earlier work by Engelhard and Behizadeh (2012) that explored consensus definitions of validity. This research has implications for improving research, theory, and practice in writing assessment.

Keywords: validity, consequential validity, writing assessment, communities of practice, Rasch measurement theory

> La validación fue alguna vez un misterio sacerdotal, un ritual tras bambalinas en el cual la élite profesional oficiaba como testigo y juez. Hoy es un espectáculo público que combina las atracciones del ajedrez y la lucha en el barro (Cronbach, 1988, p. 3).

La validez fue una idea discutida a lo largo del siglo XX, y este debate ha continuado en el siglo XXI. Desde la posición de Thorndike (1919) de que una escala válida es aquella «sobre cuyo significado coinciden todos los pensadores competentes» (p. 11) a la concepción unitaria de validez de Messick (1995) que engloba el uso planificado y las consecuencias no planificadas del uso de los resultados de un test, la comunidad de investigación en medición ha buscado un consenso con respecto al significado de la validez (Engelhard & Behizadeh, 2012; Newton, 2012). Para complicar aún más las cosas, varias comunidades específicas (por ejemplo, los investigadores en escritura) frecuentemente manejan diferentes conceptos explícitos e implícitos de validez. A medida que la validación va dejando de ser un «misterio sacerdotal» (Cronbach, 1988, p. 3) y se vuelve más un debate público, otras voces han comenzado a entrar a la discusión para contribuir a la próxima iteración de una definición de validez en evaluación. En el presente artículo, nos concentramos en evaluaciones de escritura a gran escala para escuelas primarias y secundarias, para así poder contextualizar nuestra perspectiva sobre la validez.

Subyace a nuestro estudio la idea de las *comunidades de práctica* (Wenger, 1998, 2010, 2015). Wenger (2015) identificó tres dimensiones claves de las comunidades de práctica: dominio de interés, actividades conjuntas y prácticas compartidas. Las comunidades de práctica tienen un mismo *dominio de interés* y un compromiso y competencia compartidos en este dominio. Los miembros de la comunidad se involucran en *actividades conjuntas* y discusiones que construyen relaciones que les permiten aprender los unos de los otros. Finalmente, los miembros de una comunidad de práctica tienen *prácticas compartidas* que incluyen formas de ver y de hacer frente a los problemas. Las comunidades de investigación en escritura y medición son dos comunidades de práctica que comparten un dominio de interés: la validez de los procesos de evaluación de la escritura; sin embargo, poseen conceptualizaciones distintas de la validez en la evaluación de la escritura. Históricamente, estas comunidades no se han involucrado en actividades y prácticas compartidas relativas a la validación (Behizadeh & Engelhard, 2011). Es importante apuntar que la investigación en escritura, también conocida como investigación sobre la composición, es un campo amplio e interdisciplinario que incluye a investigadores abocados a múltiples disciplinas. En este estudio, la comunidad de investigación en escritura incluye principalmente a investigadores del área del currículo y la enseñanza que se interesan en las prácticas de alfabetización, particularmente en la enseñanza y la evaluación de la escritura.

El objetivo del presente estudio es explorar definiciones del concepto de validez dentro de dos comunidades de práctica académica: la de investigación en escritura y la de medición educacional. Nos

referimos a cómo se ha presentado la validez dentro y entre las comunidades de investigación en escritura y medición. Tres preguntas claves guían nuestro estudio:

1. ¿En qué consiste una evaluación válida de escritura según la comunidad de investigación en escritura?
2. ¿En qué consiste una evaluación válida de escritura según la comunidad de investigación en medición?
3. ¿Cuáles son algunos puntos de consenso y desacuerdo sobre el concepto de validez en ambas comunidades?

Nuestras metas son fomentar la comunicación clara entre estas dos comunidades y promover avances en las áreas de la investigación, la teoría y la práctica relacionadas con la evaluación de la escritura. La teoría de medición Rasch se describe brevemente como una forma de explorar la validez dentro del contexto de la evaluación de la escritura.

Metodología

Nuestra metodología es guiada por el concepto de iteraciones epistémicas, el cual emplea análisis históricos y filosóficos para estudiar cómo los conceptos, como por ejemplo la validez, se conceptualizan y evolucionan a través del tiempo (Engelhard & Behizadeh, 2012). Nuestra revisión de estudios previos es selectiva más que exhaustiva, con el objetivo de presentar lo que consideramos investigación altamente influyente y prometedora sobre el concepto de validez.

En primer lugar, nos enfocamos en caracterizar los conceptos de validez en general presentes en las comunidades de investigación en escritura y medición. Nuestras primeras fuentes dentro de ambas comunidades son las organizaciones profesionales que entregan estándares para prácticas de evaluación. En la comunidad de investigación en escritura, seleccionamos al Consejo Nacional de Profesores de Inglés (National Council of Teachers of English [NCTE]), mientras que en la comunidad de medición elegimos al Consejo Nacional de Medición en Educación (National Council on Measurement in Education [NCME]). Se escogieron estos dos organismos debido a que nuestro estudio busca centrarse en evaluaciones de escritura a gran escala realizadas en la educación primaria y secundaria en EE.UU. El NCTE es una organización profesional de primer nivel enfocada en la enseñanza de la lengua inglesa, previa al nivel universitario, mientras que el NCME es una agrupación estadounidense de avanzada que se aboca al estudio de la medición. Examinamos los estándares de evaluación del NCTE (International Reading Association [IRA], & National Council of Teachers of English [NCTE], 2010), así como los recientemente revisados *Estándares para Tests* [Test Standards] (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Estos estándares reflejan definiciones consensuadas para cada comunidad. Luego de explorar estos estándares, seleccionamos una revista científica de primer nivel publicada por el NCTE y otra por el NCME para explorar de modo más granular la definición y el uso del término *validez* en relación con la evaluación de la escritura, centrándonos en los últimos 15 años.

En la comunidad de investigación en escritura, el término específico de *validez* no se usa al mismo nivel que en la comunidad de investigación en medición. Para explorar el uso del concepto de validez en la investigación sobre escritura y representar a la comunidad abocada a ella, elegimos la revista *Research in the Teaching of English* (RTE) del NCTE, con el mayor factor de impacto. En primer lugar, buscamos los términos «validity» o «valid*» en los títulos de artículos entre 1999 y 2014, lo que no arrojó resultados. Durante el mismo período, buscamos el término «assess*» en los títulos, con lo que obtuvimos siete resultados, tres de los cuales eran estudios empíricos relevantes para la evaluación de la escritura: Broad (2000), Bauer y Garcia (2002) y Ketter y Pool (2001). Otra búsqueda del término «valid*» en cualquier parte del texto de los artículos durante el mismo período, arrojó cinco resultados. Ketter y Pool (2001) apareció nuevamente, aunque acompañado por cuatro nuevos artículos, dos de los cuales eran relevantes para la evaluación de la escritura: Elliot, Deess, Rudniy y Joshi (2012) y Murphy (2007). Pusimos atención adicional a un número especial de RTE sobre evaluación de la escritura, publicado en 2014. En particular, seleccionamos artículos escritos por Poe (2014) y Slomp, Corrigan y Sugimoto (2014), para una revisión más exhaustiva.

En la comunidad de investigación en medición, el término *validez* se emplea con mucha frecuencia, por lo que nuestra metodología fue ligeramente distinta. Realizamos una búsqueda de términos claves, y debido a la enorme cantidad de artículos, seleccionamos investigadores del área de la medición que representan posturas emergentes claves en el debate actual sobre validez. Como en el caso de la comunidad de investigación en escritura, una revista de primer nivel (*Journal of Educational Measurement*) recientemente publicó un número especial sobre validez, que nos sirvió de guía con respecto a los puntos de vista actuales. En particular, examinamos las posturas de Kane (2013), Borsboom y colegas (Borsboom, 2005; Borsboom, Mellenbergh, & Van Heerden, 2004; Borsboom & Markus, 2013) y Sireci (2013).

En resumen, identificamos trabajos influyentes dentro de cada comunidad de práctica relativos a la discusión sobre validez. Luego, leímos cuidadosamente estos materiales seleccionados para sentar las bases de nuestra discusión detallada del concepto de validez en la evaluación de la escritura. Nuestros hallazgos destacan áreas de acuerdo y desacuerdo y tienen importantes implicaciones para la investigación, la teoría y la práctica.

¿En qué consiste una evaluación válida de la escritura desde la perspectiva de la comunidad de investigación en escritura?

En estudios anteriores, detallamos cómo evolucionó el constructo de escritura dentro de la comunidad de investigación en escritura de EE.UU., desde una perspectiva mecánica a comienzos de la década de 1900, a una definición orientada al contenido hacia mediados de siglo y, más recientemente, a definiciones socioculturales (Behizadeh & Engelhard, 2011) que posicionan al contexto como un factor importante y necesario de considerar en la enseñanza y aprendizaje de la lectoescritura. El siguiente análisis de la definición consensuada de la validez en la comunidad de investigación en escritura, representada por los *Estándares para la Evaluación de la Lectura y la Escritura* [Standards for the Assessment of Reading and Writing] (IRA/NCTE, 2010), revela el predominio de la teoría sociocultural.

Comunidad de investigación en escritura: ¿Qué es la validez?

Los *Estándares para la Evaluación de la Lectura y la Escritura* de la Organización Internacional de la Lectura y el Consejo Nacional de Profesores de Inglés (IRA/NCTE, 2010) indican que:

> Históricamente, una medición válida se ha definido de manera común como aquella que mide el constructo que pretende medir. Esto se llama *validez de constructo.* Por ejemplo, si decimos que una evaluación mide fluidez de lectura, pero en realidad solo mide velocidad y precisión y no incluye aspectos como la entonación, la prueba tendrá validez de constructo pobre (pp. 52-53).

Esta visión sobre la validez de constructo nos recuerda a nociones anteriores dentro del ámbito de la medición que empleaban una definición simple de este concepto (Kelley, 1927). Sin embargo, estos estándares continúan apuntando lo siguiente:

> Las nociones de validez más recientes incluyen un examen de las consecuencias de las prácticas de evaluación: la *validez consecuencial.* Por ejemplo, una prueba podría tener una excelente validez de constructo como medición de la habilidad de decodificación. Sin embargo, si se empleara como base para ajustar los sueldos de los profesores, causando un énfasis curricular exagerado en la decodificación, no sería un proceso de evaluación válido. En otras palabras, un procedimiento de evaluación válido no puede tener al mismo tiempo consecuencias negativas o equivocadas para los niños. En consecuencia, una definición productiva de una práctica de evaluación válida debería reflejar y apoyar el currículo valorado (IRA/NCTE, 2010, p. 53).

De acuerdo a estas definiciones, la validez de constructo y la validez consecuencial están combinadas. Parece existir un fuerte énfasis en evaluar el uso apropiado del test y las consecuencias reales de dicho uso. Un aspecto importante para la comunidad de investigación en escritura es la existencia de un propósito compartido dentro de ella, el cual guía su noción de validez: mejorar la enseñanza de los estudiantes en la asignatura de la lengua inglesa.

Además de las definiciones, es importante considerar elementos específicos de los estándares de evaluación IRA/NCTE (2010). En particular, el Estándar 7 indica que «Las consecuencias de un procedimiento de evaluación son el primer y más importante aspecto a tener en cuenta al establecer la validez de la evaluación» (p. 22). Los autores continúan apuntando que «Este estándar rechaza el argumento, desafortunadamente muy común, de que un cierto test es válido a pesar de que usarlo conlleve consecuencias problemáticas (por ejemplo, ubicar a un estudiante en un programa que no lo beneficia)» (p. 23). Además de subrayar muy claramente la importancia de la validez consecuencial, el Estándar 7 es mencionado en múltiples ocasiones en los demás estándares y también en la explicación que sigue al Estándar 1. El Estándar 1 establece que «Los intereses del estudiante son lo primordial en la evaluación» y la línea siguiente añade que «Las experiencias de evaluación en todos los niveles, ya sean formativas o sumativas, tienen consecuencias para los estudiantes (véase Estándar 7)» (p. 11).

La Figura 1 representa los cuatro temas principales de los estándares de evaluación IRA/NCTE para la lectura y la escritura. Los Estándares 3 y 6 se enfocan en el propósito de la evaluación, que consiste en informar y mejorar la enseñanza para los estudiantes lingüística y culturalmente diversos. Otro tema es el constructo de la alfabetización, representado por el Estándar 5, que articula una comprensión sociocultural de la misma, y el Estándar 4, el cual enfatiza la complejidad tanto de la lectura como de la escritura. El proceso es un tercer tema; los Estándares 2, 8, 9, 10 y 11 abogan por la inclusión de múltiples perspectivas en el proceso de evaluación de la escritura desde el desarrollo hasta el reporte, fomentando la participación activa de profesores, estudiantes, familias, administradores, legisladores y el público. El tema final es el de las consecuencias, representado por los Estándares 1 y 7, los cuales se centran en las implicaciones sociales de la evaluación para los estudiantes.

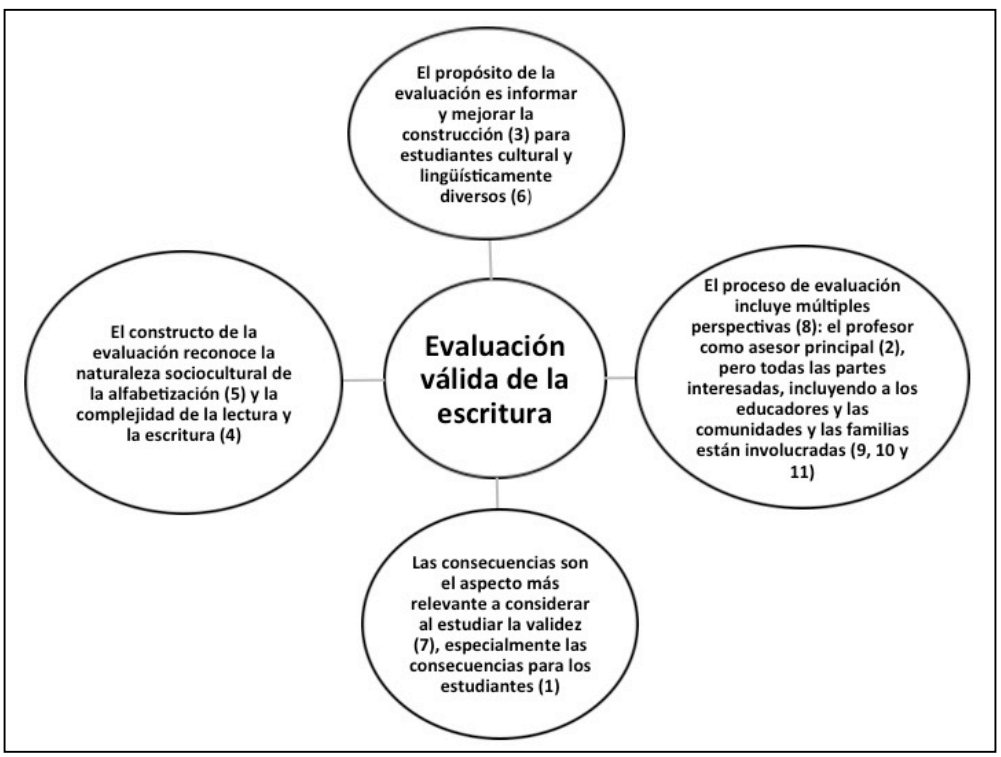

Figura 1. Comunidad de escritura: Cuatro grupos temáticos (IRA/NCTE, 2010) para una evaluación válida de la escritura (los números entre paréntesis se refieren al estándar).

Una pregunta que surge de estos estándares de evaluación es: ¿por qué se centran en la enseñanza de estudiantes diversos, en la participación de estudiantes, profesores y miembros de la comunidad y en los efectos de la evaluación sobre estos? Una forma de explicar la presencia de estos focos es que la teoría sociocultural es el paradigma dominante en la investigación sobre escritura (Behizadeh & Engelhard, 2011; Perry, 2012; Prior, 2006). La teoría sociocultural se centra en los contextos complejos y culturalmente conectados en los que ocurre la escritura y en cómo difiere esta de acuerdo al contexto. Si

una evaluación no toma en consideración la cultura de un estudiante y el contexto en el cual está aprendiendo, dicha evaluación tendrá una pobre validez de constructo, lo que conllevará consecuencias negativas no deseadas. Por ejemplo, para comprender los logros lingüísticos de un estudiante, un sistema de evaluación válido debería tomar en cuenta el contexto y permitir múltiples formas de escritura con múltiples propósitos. Esto sugiere que las evaluaciones de portafolio pueden aumentar la validez en comparación con evaluaciones de escritura directas y por demanda (Behizadeh, 2014). Los estándares de evaluación IRA/NCTE enfatizan la alta validez teórica del uso de portafolios en la evaluación sumativa debido al potencial de este método de promover prácticas de enseñanza sólidas.

Análisis de Research in the Teaching of English (RTE), 1999-2014

Sobre la base de la definición consensuada de validez materializada en los estándares IRA/NCTE (2010), ahora nos centramos en cómo emplean este término los investigadores del campo de la escritura. Escogimos un enfoque de estudio de caso para examinar cómo se utiliza la validez en la práctica dentro de esta comunidad, de modo de apoyar y/o complejizar la noción de validez presente en los estándares. En primer lugar, analizamos cinco artículos clave: Broad (2000), Ketter y Pool (2001), Bauer y Garcia (2002), Murphy (2007) y Elliot et al. (2012).

El artículo de Broad (2000) describe las prácticas de puntuación de profesores involucrados en un programa universitario de composición en inglés para estudiantes de primer año. Una evaluación válida de la escritura, de acuerdo a Broad (2000), toma en consideración la función retórica de esta, la cual no puede ser estandarizada. Para Broad, la pregunta central de la validez es: «¿Cómo podemos alinear nuestras evaluaciones con nuestras teorías y pedagogías sobre la retórica?» (p. 215). Así, podemos observar el predominio de la validez de constructo cuando el constructo de la escritura es dependiente del contexto. Con respecto a las consecuencias, Broad (2000) asevera que

> Cronbach (1990) y Messick (1989) argumentaron que la *validez consecuencial* —es decir, el análisis del impacto institucional y social de un programa de evaluación dado— es tan importante como la validez de constructo o la validez predictiva. Puesto que la evaluación inevitablemente impulsa la enseñanza y el aprendizaje, evaluaciones auténticas como… los portafolios de escritura reflejan de mejor manera la complejidad y el contexto dentro del cual tuvo lugar el aprendizaje y entregan validez consecuencial al apoyar las buenas prácticas en la pedagogía de la composición (pp. 250-251).

De manera similar a los estándares IRA/NCTE, Broad (2000) enfatiza tanto la validez de constructo como la validez consecuencial.

En el estudio de Bauer y Garcia (2002) se evalúa la validez de contenido (término hasta cierto punto intercambiable con el de validez de constructo dentro de la comunidad de investigación en escritura) y se presenta el *efecto colateral* positivo de un portafolio de alfabetización en la sala de clases. Con respecto a la validez de contenido y las evaluaciones alternativas, Bauer y Garcia apuntan: «Este tipo de evaluaciones refinan el tema de la validez de contenido al evaluar el desempeño real de los estudiantes en tareas relacionadas con un dominio» (p. 464). Los autores contrastan la autenticidad hipotética de las evaluaciones alternativas, como los portafolios, con la de las pruebas estandarizadas, las cuales «en el mejor de los casos, funcionan como una representación indirecta del desempeño» (p. 464). En la sección siguiente, Bauer y Garcia (2002) se refieren inmediatamente a la validez consecuencial, indicando que:

> Dado que los profesores en contextos de alto impacto tienden a enseñar para la prueba (Center for the Study of Testing, Evaluation, & Educational Policy, 1992), varios investigadores han sugerido que las pruebas alternativas, apoyadas por estándares públicos, podrían producir mayor equidad para los estudiantes, especialmente en el caso de aquellos con bajos ingresos y bajo nivel de desempeño (p. 464).

Los autores continúan explicando que las evaluaciones alternativas podrían aumentar la equidad al permitir que los estudiantes accedan a enseñanza y a evaluaciones de alta calidad, además de darle mayor relevancia a la voz de los estudiantes en el proceso educativo.

De acuerdo a los dos artículos revisados hasta ahora, las consecuencias de las evaluaciones de escritura son el aspecto primordial. En este sentido, Broad (2000) y Bauer y Garcia (2002) promueven formas de evaluación de la lectoescritura basados en portafolios debido al impacto positivo de este método sobre la enseñanza. Aunque existen otros modos de conceptualizar la validez consecuencial, la comunidad de investigación en escritura parece enfocarse en el *efecto colateral* de la evaluación a la enseñanza.

Complementando el argumento de Bauer y Garcia (2002) con respecto a que las evaluaciones alternativas centradas en el estudiante aumentan la equidad de la enseñanza, Ketter y Pool (2001) exploraron el impacto negativo de las pruebas de escritura de alto impacto en la enseñanza. Específicamente, examinaron dos aulas a las que asistían estudiantes secundarios con problemas de escritura y observaron cómo una prueba de escritura de alto impacto afectó la enseñanza. Además de citar a Messick (1989) y a Cronbach (1990), Ketter y Pool (2001) citan el trabajo de Moss (1994), destacando especialmente su expansión del concepto de validez para incluir la validez consecuencial. A lo largo de su artículo, los autores enfatizan las consecuencias de las pruebas de escritura estandarizadas de alto impacto, específicamente, los efectos negativos de estas pruebas sobre la enseñanza. Nuevamente, se hace énfasis en la validez consecuencial, tal como en los otros artículos analizados en este estudio y en los estándares IRA/NCTE (2010).

Aunque Ketter y Pool (2001) destacan la validez consecuencial en su trabajo, como otros investigadores del campo de la escritura, notan la conexión entre esta y la validez de constructo:

> Debido a la naturaleza especializada de muchas conversaciones sobre evaluación de la escritura en la comunidad de investigación en medición, los investigadores del campo de la composición consideran que las justificaciones para realizar evaluaciones directas de esta son innecesariamente técnicas y finalmente irrelevantes, puesto que ignoran un vasto conjunto de literatura especializada que teoriza sobre cómo las personas aprenden a leer y escribir y que pone en duda la visión de la habilidad de escritura como un rasgo fijo que no es afectado por el contexto (p. 347).

Incluso cuando los estudios se enfocan en la validez consecuencial, como en el caso del de Ketter y Pool, un punto esencial consiste en establecer cómo estos tests definen la escritura.

En comparación con los tres artículos anteriores, Elliot et al. (2012) y Murphy (2007) presentan diferencias significativas. En primer lugar, el artículo de Murphy (2007) es un documento de posición basado en estudiantes que son minorías lingüísticas y sobre la evaluación, mientras que la investigación de Elliot et al. (2012) es un estudio cuantitativo de validación que emplea evidencia sobre validez predictiva y concurrente, en otras palabras, correlaciones entre puntuaciones en un test de nivelación y (a) las notas finales en un curso y (b) otras pruebas de escritura. Asimismo, tanto Elliot et al. (2012) como Murphy (2007) citan los *Estándares para Tests* AERA, APA y NCME (1999) en lugar de estándares creados por organizaciones especializadas en alfabetización. Elliot et al. (2012) se basan en gran medida en el trabajo de Kane (2006), un teórico de la medición, lo que sitúa a ambos estudios como puentes entre las comunidades de investigación en escritura y medición. Sin embargo, como todas las investigaciones de *RTE* revisadas en el presente estudio, los autores expresan que su preocupación con respecto a la validez de constructo constituye el centro de su investigación. Para explicar por qué eligieron comparar las puntuaciones de pruebas de nivelación con las de portafolios de fin de curso, Elliot et al. (2012) expresan que «la representación limitada del constructo de la escritura en pruebas compradas y la manifestación de dicho constructo en los portafolios son aspectos muy significativos para nosotros» (p. 290). Asimismo, Murphy (2007) indica que el campo del lenguaje y la alfabetización ha progresado en su comprensión de la lectoescritura como algo sociocultural y situado, mientras que algunos investigadores del campo de la medición siguen definiendo la lectura y la escritura como habilidades discretas que pueden existir ajenas a consideraciones lingüísticas o culturales. En términos de validez consecuencial, Elliot et al. (2012) discuten cómo la atención que prestaban a la validez predictiva se basaba en su preocupación por las consecuencias sociales de emplear esta prueba de nivelación, mientras que Murphy (2007) dedica varias páginas a detallar el disímil impacto de las pruebas de lenguaje en estudiantes cultural y lingüísticamente diferentes, citando la idea de «validez cultural» como una fuente necesaria de evidencia para la validez.

En suma, Broad (2000), Ketter y Pool (2001), Bauer y Garcia (2002) y Murphy (2007) reflejan la visión de los investigadores del campo de la escritura con respecto a la evaluación, prestando especial

atención a la validez consecuencial por sobre todas las demás preocupaciones. De modo similar, Elliot et al. (2012) destacan la validez consecuencial como un asunto clave, aunque ellos emplean una metodología distinta. En todos los artículos, la falta de validez de contenido se posiciona como la causa de las consecuencias no planificadas. Estos artículos clave están bien alineados con el Estándar 1 IRA/NCTE (2010): «Las consecuencias de un procedimiento de evaluación son el primer y más importante aspecto a tener en cuenta al establecer la validez de la evaluación» (p. 22). Antes de pasar al número especial de *RTE,* es importante considerar que estos artículos son, en cierto grado, casos atípicos, porque históricamente la comunidad de la escritura se ha enfocado más en las prácticas de enseñanza que en las de evaluación. Esto se relaciona con la cita de Cronbach (1988) al comienzo del presente artículo: la evaluación puede haber sido percibida por los investigadores del campo de la escritura como un feudo de la comunidad de la medición. Solo en el último tiempo la comunidad de investigación en escritura ha comenzado a abocarse a la tarea de explorar problemas relativos a la validez.

Número especial de Research in the Teaching of English (RTE)

Un ejemplo del creciente grado de atención que se le presta a la validez en la comunidad de investigación en escritura es el reciente número especial de *RTE* sobre la evaluación de la escritura. Estableciendo el claro enfoque de esta serie de artículos, Poe (2014) titula su introducción al volumen (2014) «The Consequences of Writing Assessment» [«Las consecuencias de la evaluación de la escritura»]. En esta introducción, Poe (2014) destaca el desarrollo histórico de nuevas revistas diseñadas para ser puentes entre la investigación sobre la escritura y la investigación sobre medición, refiriéndose a dos casos en particular: *Language Testing,* publicada por primera vez en 1984, y *Assessing Writing,* lanzada en 1994. Poe (2014) articula las metas de estas revistas del siguiente modo:

> El deseo de reconocer múltiples perspectivas, incluyendo las de los profesores, impulsó el progreso teórico. Sin embargo, a pesar de los nuevos espacios para reconocer múltiples perspectivas en la evaluación de la escritura, otros avances en este campo, como los modelos socio-cognitivos de la escritura, fueron pasados por alto y los efectos locales de la evaluación de la escritura sobre aprendices diversos fueron frecuentemente ignorados (p. 272).

Como en artículos anteriormente publicados en *RTE,* Poe (2014) se enfoca en la validez consecuencial. Asimismo, Poe (2014) cita a Kane (2006), continuando la tendencia establecida a partir de Ketter y Pool (2001), Broad (2000) y Elliot et al. (2012). Después de hacer referencia a Kane, Poe (2014) expone que «los investigadores hoy en día se enfocan en cómo debe reinterpretarse el constructo de la escritura para cumplir con los requerimientos de la evaluación localizada de la escritura» (p. 273). Esta cita se refiere tácitamente a la importancia de la validez de constructo, aunque la autora no entrega detalles específicos sobre cómo los investigadores del área de la educación debieran definir la escritura.

Hasta cierto punto, todos los artículos de este número especial de *RTE* destacan la importancia de la validez consecuencial. En particular, el artículo de Slomp et al. (2014) titulado «A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study» [«Marco para emplear evidencia sobre validez consecuencial en evaluaciones de escritura a gran escala: un estudio canadiense»] remarca este interés constante en las consecuencias. Los autores citan a Messick (1989) y a Cronbach (1988) para apoyar su decisión de enfocarse en la validez consecuencial, y luego argumentan que los modelos de validez que se centran en la validez de constructo no son suficientes para examinar la validez consecuencial, «porque las consecuencias de los tests y sus interpretaciones muchas veces superan los límites de la interpretación y el uso de estos» (p. 278). Los autores toman elementos del modelo de validez de Kane (2006, 2013) para desarrollar su propio marco para examinar la validez consecuencial en evaluaciones a gran escala, empleando la evaluación de la escritura en Canadá como estudio de caso. Slomp et al. (2014) concluyen su trabajo de forma muy similar a Ketter y Pool (2001), expresando lo siguiente:

> Si el verdadero objetivo de los programas de evaluación a gran escala impulsados por el gobierno es mejorar los sistemas educacionales, entonces es lógico que los diseñadores y usuarios de los tests examinen y reporten públicamente las consecuencias del empleo de estos (p. 298).

A lo largo de esta sección, los investigadores del campo de la escritura han hecho énfasis en la importancia de evaluar usos y consecuencias reales, un énfasis que requiere de investigación empírica en las aulas para examinar los efectos de las evaluaciones de la escritura sobre la enseñanza y los estudiantes.

¿En qué consiste una evaluación válida de la escritura según la comunidad de investigación en medición?

Los primeros teóricos de la medición definían una escala válida de forma aparentemente sencilla: «aquella sobre cuyo significado coinciden todos los pensadores competentes» (Thorndike, 1919, p. 11). Aunque referirnos en detalle a la historia de la evolución del concepto de validez dentro de la comunidad de la medición escapa a los límites del presente estudio, existen varias perspectivas claves que deben ser destacadas. A comienzos del siglo XX, Thorndike (1914) definió las escalas válidas esencialmente como un consenso sobre el significado de las puntuaciones dentro de una comunidad de profesionales. Años más tarde, Kelly (1927) definió los tests válidos en términos de un consenso respecto a lo que *pretenden* medir las escalas. Tanto Thurstone (1931) como Gulliksen (1950) sugirieron que la validez de los tests fuera evaluada utilizando evidencia relacionada con criterios y basada en datos empíricos. Cronbach (1971) redireccionó el relato acerca de la validez desde las escalas válidas a la *validación de los tests,* la que se define como un proceso de evaluación basado en el propósito planificado de las puntuaciones de los tests y en sus usos recomendados. En su trabajo pionero, Messick (1989) propuso una perspectiva amplia con respecto a la validez de constructo la cual consideraba tomar en cuenta la evidencia sobre validez consecuencial. Más recientemente, Kane (2013) ha promovido un enfoque basado en argumentos para estudiar la validez (véase Engelhard & Behizadeh, 2012 para obtener más detalles sobre las iteraciones históricas y filosóficas del concepto de validez).

A pesar de que esta es una visión ampliamente aceptada de la evolución del concepto de validez, la comunidad de investigación en medición sigue en desacuerdo con respecto a ciertos aspectos del término (Engelhard & Behizadeh, 2012; Newton, 2012). En la sección siguiente, describimos la definición consensuada de validez encarnada en los *Estándares para Tests* (AERA et al., 2014). Luego, empleamos un reciente número especial de *Journal of Educational Measurement* para representar la discusión actual sobre validez dentro de la comunidad de investigación en medición antes de comparar los puntos de vista sobre validez entre las comunidades. Finalmente, ilustramos cómo los avances en medición representados por la teoría de medición Rasch (Engelhard, 2013; Rasch, 1960/1980) pueden ser utilizados para entregar una visión coherente de la evidencia sobre validez, tanto en general como específicamente dentro del contexto de la evaluación de la escritura.

Comunidad de investigación en medición: ¿Qué es validez?

> La validez se refiere al grado en el cual la evidencia y la teoría apoyan las interpretaciones de las puntuaciones de los tests para los usos propuestos de estos (AERA et al., 2014, p. 11).

La definición consensuada actual de validez, delineada en los recientemente revisados *Estándares para Tests* (AERA et al., 2014) es la siguiente: «Debe establecerse una articulación clara de cada propósito de interpretación de las puntuaciones para un uso específico, y debe proveerse evidencia apropiada sobre validez para apoyar cada interpretación planificada» (AERA et al., 2014, p. 23). Este principio rector se divide en tres grupos temáticos:

I. Establecer usos e interpretaciones planificados,
II. Problemas con respecto a muestras y contextos usados en la validación y
III. Formas específicas de evidencia sobre validez.

En la Figura 2 se muestran estos tres grupos. La figura destaca la idea de que la validez no puede evaluarse en el vacío, sin considerar específicamente los usos e interpretaciones planificados (Grupo I) ni el contexto del sistema de evaluación relacionado con las personas y el entorno involucrados en el proceso de validación (Grupo II). Sin embargo, muchas discusiones sobre validez se enfocan principalmente (y muchas veces exclusivamente) en el último grupo (Grupo III). Hasta cierto punto, tanto la comunidad de investigación en medición como la comunidad de investigación en escritura

tienden a enfocarse en formas específicas de evidencia sobre validez, como la de contenido, la de constructo, la basada en criterios y la consecuencial sin referirse explícitamente a los propósitos de las interpretaciones. Se recomienda a los lectores consultar los *Estándares para Tests* para obtener una descripción más acabada de cada grupo.

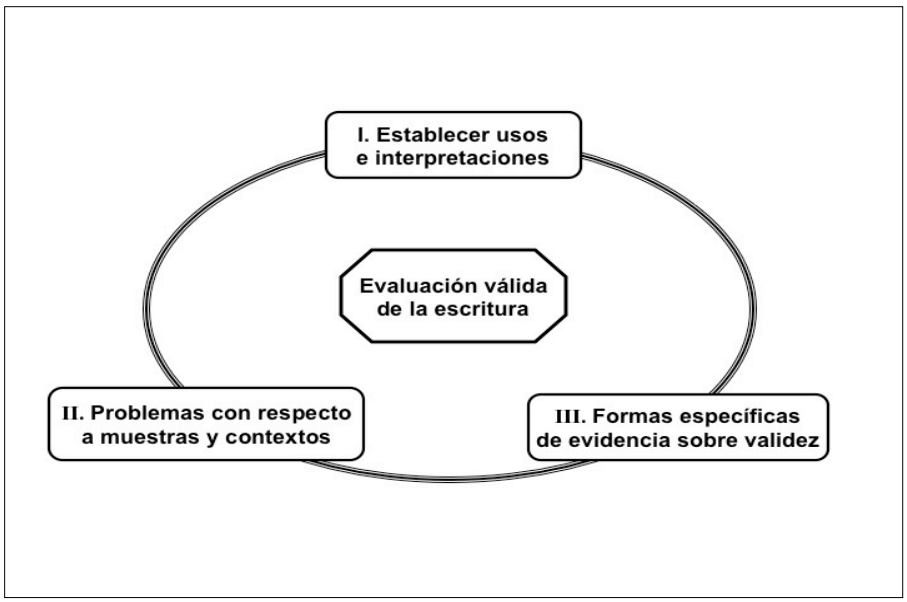

Figura 2. Comunidad de investigación en medición: tres bloques temáticos (AERA et al., 2014).

Establecer el propósito e interpretaciones de un sistema de evaluación es esencial para establecer la validez de la interpretación y del uso de las puntuaciones de los tests desde el punto de vista de los *Estándares para Tests.* Los tests en el área de la educación, incluyendo las evaluaciones de escritura, están diseñadas para cumplir variados propósitos. A pesar de la importancia de establecer el objetivo del sistema de evaluación, es sorprendente la poca claridad que tienen los propósitos de muchos tests educacionales, incluyendo las evaluaciones de lectura, a los ojos de muchas partes interesadas.

El segundo grupo temático subraya que los asuntos relativos a la validación son dependientes del contexto, y que debería entregarse la información lo más detallada posible sobre los participantes del sistema de evaluación. Esto incluye las características demográficas de las muestras y los contextos. Este énfasis en el contexto es hasta cierto punto congruente con una perspectiva sociocultural de la escritura, aunque muchos miembros de la comunidad de medición podrían no estar completamente conscientes de todo lo que implica incluir este grupo temático al presentar evidencia sobre validez. Los estándares IRA/NCTE (2010) indican que una evaluación verdaderamente sociocultural debería involucrar a todas las partes interesadas, incluyendo a los estudiantes y a las familias, en el proceso completo de validación, desde el desarrollo hasta el informe final. Esta es un área donde la colaboración entre comunidades tiene el potencial de generar mejoras significativas en nuestra comprensión conceptual de las consecuencias planificadas y no planificadas de las evaluaciones de escritura.

El último grupo temático incluye seis formas de evidencia sobre validez:

- Evidencia orientada al contenido,
- Evidencia con respecto a procesos cognitivos,
- Evidencia con respecto a estructura interna,
- Evidencia con respecto a relaciones con constructos conceptualmente asociados,
- Evidencia con respecto a relaciones con criterios y,
- Evidencia basada en las consecuencias de los tests.

Los *Estándares para Tests* (AERA et al., 2014) destacan la importancia del propósito de todo sistema de evaluación, así como la recolección de evidencia sobre validez para justificar los propósitos e

interpretaciones de las puntuaciones de los tests. También es relevante considerar las consecuencias planificadas y no planificadas de los sistemas de evaluación (Engelhard & Wind, 2013).

Número especial sobre validez: Journal of Educational Measurement (JEM)

Una de las principales organizaciones profesionales del campo de la medición educacional en EE.UU. es el Consejo Nacional de Medición en Educación [National Council on Measurement in Education (NCME)], y su publicación principal es *Journal of Educational Measurement (JEM).* Un número especial recientemente publicado de *JEM* sobre validez incluye un texto de Kane (2013), el cual refleja una actualización de sus posturas sobre un enfoque de la validación basado en argumentos (Kane, 1992, 2006), y además cuenta con aportes de Sireci (2013) y Borsboom y Markus (2013). La perspectiva basada en argumentos de Kane (2013) es congruente con posturas anteriores de Cronbach (1988):

> Idealmente, los validadores se prepararán como los participantes de un debate… planificando argumentos en favor y en contra de tan buena forma que podrían defender cualquiera de las dos posturas. O, llevando la metáfora a los consejeros legales, prepararse tan bien como para decirle a cualquiera de los litigantes cuáles son los puntos fuertes y débiles en su postura (p. 3).

Contrastando con esta visión sobre la validez basada en argumentos, Kane (2013) se refiere a Borsboom y colegas, quienes propusieron un enfoque sobre la validación basado en atributos, rasgos y constructos (Borsboom, 2005; Borsboom et al., 2004; Borsboom & Markus, 2013). Estos investigadores indican que «un *test es válido* para medir un atributo sí y solo sí (a) el atributo existe y (b) las variaciones en el atributo producen variaciones en los resultados del procedimiento de medición de modo causal» (Borsboom et al., 2004, p. 150), una idea que encuentra eco en Borsboom y Markus (2013), también parte del número especial. Esta visión de la validez basada en atributos tiene distintas implicaciones para definir el proceso de validación usado para valorar un sistema de evaluación. Por ejemplo:

> Lo que necesita probarse no es una teoría sobre la relación entre el atributo medido y otros atributos, sino que una teoría de comportamiento de respuesta. En algún punto de la cadena de eventos que ocurre entre la administración del ítem y la respuesta al ítem, el atributo medido debe tener un rol causal en determinar el valor que tendrán los resultados de las mediciones; en caso contrario, el test no puede ser válido para medir el atributo (Borsboom et al., 2004, p. 1062).

Este acercamiento a la validez basado en atributos está alineado con visiones anteriores sobre la validez que consideraban a los constructos como una variable latente subyacente que una evaluación está diseñada para medir. Borsboom, Mellenbergh y Van Heerden (2004) reposicionan al test, más que a su interpretación o su uso, cómo válido, lo cual contrasta con el enfoque basado en argumentos que actualmente predomina en los *Estándares para Tests* (AERA et al., 2014), el cual parece haber alejado a la comunidad de investigación en medición de una visión sobre la validez impulsada por los atributos y los constructos.

Por otra parte, Sireci (2013) ofrece lo que consideramos una prometedora reconceptualización del proceso de validación que se conecta estrechamente con el grupo tres de los *Estándares para Tests.* En sus comentarios sobre Kane (2013), Sireci (2013) indica que:

> Una declaración explícita de los objetivos de la evaluación es el inicio lógico de la validación, pero yo voy un paso más adelante: es el comienzo lógico del desarrollo de un test. Es decir, los tests son desarrollados para cumplir uno o más propósitos determinados. Ayudar a las personas que encargan estos tests a articular dichos propósitos depende de quienes trabajamos en el ámbito de la psicometría. Una vez que se articulan los propósitos, sabemos qué es lo que necesitamos validar. ¡Y también sabemos qué es lo que necesitamos medir! (p. 100).

Sireci (2013) identifica tres pasos a lo que él llama «plan de validación», que son: (a) articular claramente los objetivos, (b) tener en cuenta los potenciales malos usos del test y (c) cruzar los objetivos con la recolección de evidencia sobre validez. Estos pasos se basan directamente en el proceso de validación más complicado de Kane (2013), aunque Sireci combina la interpretación y el uso en un solo

término: propósito. Las recomendaciones de Sireci (2013) pueden ser aplicadas al contexto específico de la evaluación de la escritura para así determinar qué formas de validez se requieren.

¿Cuáles son algunos puntos de consenso y de debate sobre el concepto de validez?

> La validación es una responsabilidad compartida de quien desarrolla el test y de quien lo usa (AERA et al., 2014, p. 22).

Parece existir un consenso general dentro de la comunidad de investigación en escritura con respecto a las potenciales consecuencias negativas no planificadas que pueden tener sobre la enseñanza las evaluaciones estandarizadas de escritura. La validez consecuencial es claramente una preocupación central en la comunidad de investigación en escritura, particularmente en la evaluación entre kindergarten y el cuarto año medio, de acuerdo a lo que se representa en la teoría, los estándares y la investigación empírica. Por otra parte, la reciente publicación de los *Estándares para Tests* (AERA et al., 2014) entrega un marco consensuado para definir la validez, pero esta publicación no captura completamente los matices del persistente debate sobre este concepto fundacional, tal como lo evidencia el número especial sobre validez en *Journal of Educational Measurement* (2013). Más aún, los *Estándares para Tests* no desarrollan completamente cómo deberían implementarse estas orientaciones en distintas áreas de evaluación, como la evaluación de la escritura. Es bastante común que buena parte de la investigación psicométrica sobre la evaluación de la escritura se centre únicamente en los índices de acuerdo entre observadores, sin una perspectiva más amplia de validez como la que se promueve en la comunidad de investigación en medición.

Volviendo a los estándares de ambas comunidades, creemos que la siguiente visión sobre la validez recogida de los *Estándares para Tests* podría ser atractiva para los miembros de ambas comunidades:

> Finalmente, la validez de una interpretación determinada de las puntuaciones de un test descansa sobre toda la evidencia disponible relativa a la calidad técnica de un sistema de evaluación. Los capítulos siguientes de los *Estándares* describen diferentes componentes de la evidencia sobre la validez, e incluyen evidencia sobre la cuidadosa construcción de los tests; la confiabilidad adecuada de las puntuaciones; la administración y puntuación apropiada de los tests; y la atención especial en que los tests sean justos para todos quienes los rindan, de acuerdo a la interpretación del test en cuestión (AERA et al., 2014, p. 22).

Esta idea de validez está alineada hasta cierto punto con los estándares IRA/NCTE (2010), con un claro énfasis en el examinado como una parte interesada clave en el proceso de evaluación. Sin embargo, pueden existir grandes diferencias entre cómo los investigadores del campo de la medición y de la escritura conceptualizan qué es lo que significa desarrollar *cuidadosamente* una prueba, cómo se define la confiabilidad, en qué consiste la administración *apropiada* y, lo que tal vez sea lo más importante, qué quiere decir el término *justo.* Por ejemplo, si lo justo se equipara con la estandarización en la comunidad de investigación en medición (Madaus, 1994), esta postura entra en conflicto con la visión que predomina en la comunidad de investigación en escritura, lo que consiste en permitir que múltiples perspectivas tengan voz durante el todo el proceso de evaluación (IRA/NCTE, 2010).

Los investigadores de los campos de la escritura y la medición ofrecen sus propias características basadas en sus marcos teóricos y su posición con respecto al proceso de validación. Consideramos que existe un gran potencial para aumentar la colaboración entre estas comunidades, y ya hemos detectado un grado no despreciable de préstamos teóricos desde la comunidad de investigación en medición por parte de investigadores del área de la escritura. Sin embargo, esperamos que este proceso de préstamo se vuelva más recíproco, especialmente en lo que concierne a asegurar la equidad de las evaluaciones para estudiantes cultural y lingüísticamente diversos (Murphy, 2007). Específicamente, los investigadores del área de la escritura están en la mejor posición para evaluar la validez consecuencial, especialmente las consecuencias negativas no planificadas, mediante estudios empíricos llevados a cabo en colaboración con profesores de aula. Los investigadores del campo de la medición pueden entregar consejos técnicos y sugerencias para realizar estudios sobre la evaluación de la escritura, incluyendo la examinación de la invarianza de medición, como es el caso del funcionamiento diferencial de los ítems, un aspecto explorado en la próxima sección. Los *Estándares para Tests* revisados (AERA et al., 2014) dejan muy claro que la validez es una responsabilidad conjunta de quienes desarrollan las pruebas, y de quienes las

usan, y los estándares IRA/NCTE (2010) enfatizan la importancia de que todas las partes interesadas estén involucradas, incluyendo las comunidades locales donde tienen lugar las evaluaciones. Ambas comunidades de investigación deberían ver esto como una oportunidad para mejorar la teoría y la práctica de la evaluación de la escritura.

Nuestros análisis sugieren que los investigadores del campo de la escritura están interesados principalmente en los usos *reales* de la evidencia sobre validez, mientras que los investigadores del campo de la medición hacen un mayor énfasis en establecer el propósito *determinado* como fuente de evidencia sobre validez. Creemos que esta diferencia clave en la conceptualización de la validez consecuencial deriva de las diferencias en los propósitos percibidos de evaluaciones de escritura a gran escala. El marco de Sireci (2013) tiene potencial para informar las narrativas emergentes sobre validez originadas en las comunidades de investigación en escritura y medición. Adicionalmente, Slomp et al. (2014) propusieron un marco prometedor para examinar la validez consecuencial en evaluaciones de escritura a gran escala que puede ser útil para ambas comunidades. Estos dos marcos se exploran a continuación.

Tres de las recomendaciones de Slomp y colegas (2014) son particularmente relevantes para nuestra discusión. La primera se basa en la observación de que las partes interesadas más alejadas del aula veían positivamente la evaluación a gran escala, mientras que profesores y estudiantes tendían a presentar actitudes negativas con respecto a esta. Al parecer, estos desacuerdos derivan principalmente de una distinta percepción de los propósitos determinados de las evaluaciones de la escritura. En segundo lugar, Slomp et al. (2014) abogan por una mayor discusión entre las comunidades de investigación en escritura y medición, una meta que intentamos llevar a la práctica en el presente artículo. Finalmente, sugieren que se emplee su marco para recoger evidencia relacionada con las consecuencias de la evaluación.

Existe una interesante correspondencia entre Sireci (2013), con su énfasis en el propósito y en la evidencia sobre validez, y la propuesta de Slomp et al. (2014) de examinar la evidencia sobre validez consecuencial. El cruce hecho por Sireci (2013) entre los seis tipos de evidencia sobre validez y las consecuencias planificadas y no planificadas de una evaluación, ofrece una herramienta que puede ser utilizada en las comunidades de investigación en medición y en escritura para llevar a cabo un proceso de validación conjunto. Nuestra razón para elegir el plan de validación de Sireci es que incluye la exhaustiva base teórica delineada por Kane (2013), pero a la vez simplifica el proceso de validación. Esta simplificación podría permitir que esta herramienta sea más transportable entre comunidades y que ingrese más fácilmente a la comunidad de investigación en escritura.

En resumen, ambas comunidades de investigación han desarrollado herramientas conceptuales valiosas para mejorar procesos de evaluación válidos, las que podrían ser incluso más poderosas si se emplearan juntas. El marco de Sireci (2013) puede funcionar fácilmente en conjunto con la matriz de preguntas sobre validez consecuencial de Slomp et al. (2014). La Tabla 1 ilustra el cruce de la evidencia sobre validez y las consecuencias de una hipotética evaluación de escritura, empleando las formas específicas de evidencia sobre validez recomendadas por los *Estándares para Tests* actualizados (AERA et al., 2014) e ilustradas por Sireci (2013), con la adición de preguntas adaptadas de la matriz de preguntas de Slomp y colegas (2014). Si el propósito de una prueba de escritura es determinar niveles de competencia para una evaluación sumativa, sugerimos que varias preguntas de la Tabla 1 sean examinadas considerando las consecuencias planificadas y no planificadas. Recomendamos enérgicamente que las consecuencias no planificadas sean exploradas explícitamente junto con cada tipo de evidencia sobre validez, y que tanto las consecuencias hipotéticas como las reales sean investigadas y documentadas. La intención de esta tabla es servir como ilustración y no es de ningún modo exhaustiva en cuanto a tipos de evidencia, consecuencias o preguntas subyacentes sobre validez. Consideramos que la Tabla 1 constituye un punto de partida para desarrollar una división de tareas entre estas dos comunidades con respecto a áreas de especialización en teoría y metodologías. Por ejemplo, la comunidad de investigación en escritura probablemente esté en contacto con profesores encargados de enseñar a escribir a sus alumnos y puede contribuir con puntos de vista útiles sobre cómo las evaluaciones están siendo efectivamente usadas en las aulas, incluyendo la evaluación de consecuencias tanto planificadas como no planificadas.

Tabla 1
Evidencia sobre validez cruzada con consecuencias planificadas y no planificadas

Propósito de la evaluación de escritura: Determinar niveles de competencia en escritura mediante un ensayo escrito		
	Consecuencias	
Formas específicas de evidencia sobre validez	Planificadas	No planificadas
Evidencia orientada al contenido	¿Representan las puntuaciones en el test el constructo de la escritura?	¿Representan las puntuaciones en el test el constructo de la escritura para estudiantes de variados contextos socioculturales?
Evidencia con respecto a procesos cognitivos	¿Reportan los estudiantes tener los pensamientos y comportamientos esperados en relación con el constructo de la escritura?	¿Desarrollan los estudiantes conceptos errados sobre la escritura?
Evidencia con respecto a estructura interna	¿Apoyan los componentes del sistema de evaluación las inferencias hechas con respecto al constructo de la escritura?	¿Pueden replicarse los componentes del sistema de evaluación de escritura en distintos contextos socioculturales?
Evidencia con respecto a relaciones con constructos conceptualmente asociados	¿Se asocian las puntuaciones en escritura con otros constructos relacionados?	¿Se asocian las puntuaciones en escritura con otras variables irrelevantes para el constructo?
Evidencia con respecto a relaciones con criterios	¿Se asocian las puntuaciones en escritura con otros indicadores de alfabetización?	¿Se asocian las puntuaciones en escritura con otros indicadores de alfabetización para estudiantes de variados contextos socioculturales?
Evidencia basada en las consecuencias de los tests	¿Logran los estudiantes el desempeño esperado en otras tareas que requieren competencia en escritura?	¿Enseñan los profesores a rendir el test y limitan el currículo relacionado con la escritura?

Nota: Tabla basada en Sireci (2013) y Slomp et al. (2014).

Los propósitos de un sistema de evaluación definen tanto los usos directos como indirectos con consecuencias potencialmente no planificadas cada vez que se pone en marcha una evaluación de escritura dentro de un contexto complejo. Puesto que las comunidades de investigación en medición y en escritura toman parte en conversaciones sobre validez, deberían hacer un mayor esfuerzo por expresar explícitamente los propósitos de las evaluaciones de escritura y examinar las implicaciones potencialmente contradictorias de propósitos múltiples. Como se mencionó anteriormente, un tema clave que emergió fue el contraste entre uso planificado y real. Con respecto a este punto, Kane (2013), Sireci (2013) y otros investigadores del área de la medición hacen énfasis en los usos planificados y los malos usos hipotéticos o las consecuencias no planificadas, mientras que los investigadores abocados a la escritura con frecuencia exploran los usos reales y documentan las consecuencias negativas no planificadas (por ejemplo, Slomp, Corrigan, & Sugimoto, 2014). Junto con la integración del trabajo de Sireci (2013) y Slomp et al. (2014) en la Tabla 1 como una herramienta para explorar consecuencias planificadas y reales, las herramientas de la teoría de medición Rasch, detalladas en la sección siguiente, se adaptan particularmente bien a estas exploraciones.

Evidencia sobre validez desde la perspectiva de la teoría de medición Rasch

Esta sección se refiere brevemente a la teoría moderna de la medición de acuerdo a la descripción de Embretson (1996), empleando la teoría de medición Rasch (Engelhard, 2013; Rasch, 1960/1980). Los *Estándares para Tests* tienden a ser neutrales en términos de cuáles teorías de medición se emplean para entregar evidencia sobre validez para los usos planificados de las puntuaciones obtenidas de una evaluación. La teoría de medición Rasch entrega la oportunidad de examinar explícitamente las respuestas de cada estudiante a cada ítem o tarea en un sistema de evaluación, lo que entrega evidencia

sobre validez que puede usarse para evaluar el propósito (Sireci, 2013) y las consecuencias reales de las evaluaciones (Slomp et al., 2014). La meta general de los sistemas de evaluación es crear mediciones invariantes con las puntuaciones de los estudiantes independientemente de observadores y tareas específicos. En nuestro trabajo, hemos notado que es útil emplear la teoría de medición Rasch como modelo subyacente que funcione como una pieza (Engelhard, 2013) dentro del marco de medidas de construcción de Wilson (2005). El marco de Wilson puede combinarse con la teoría de medición Rasch para recoger evidencia sobre validez sistemáticamente, basándose en las seis formas de evidencia sobre validez descritas en los *Estándares para Tests* (Duckor, Draney, & Wilson, 2009) y destacadas en el trabajo de Sireci (2013) (véase Tabla 1); además, los datos cualitativos pueden emplearse para apoyar análisis cuantitativos.

Es esencial recoger evidencia orientada al contenido para fundamentar el uso de puntuaciones de tests pertenecientes a una evaluación de escritura, para determinar niveles competencia en escritura (emergente, básico, competente y avanzado). La evidencia relacionada con el contenido refleja el alineamiento entre el contenido de la evaluación y el constructo que se planea que representen las puntuaciones en el test. Sobre la base de la teoría de medición Rasch y el marco de medidas de construcción, deben considerarse múltiples pasos, incluyendo (a) la definición del constructo de escritura, (b) la descripción de las indicaciones y las tareas de escritura, (c) las reglas para puntuar los ensayos de los estudiantes (rúbricas) y (d) la creación de un mapa de variables. Wilson (2005), Duckor, Draney y Wilson (2009) y Engelhard (2013) describen en detalle cómo pueden combinarse sistemáticamente estos pasos para entregar evidencia relacionada con el contenido. Visualizamos que estos pasos se lleven a cabo como una empresa conjunta (Wenger, 2010, 2015) entre las dos comunidades, en la cual los expertos de los campos de la escritura y la medición participen en conversaciones interdisciplinarias que puedan representar las necesidades de distintas partes interesadas asociadas con la evaluación a gran escala de la escritura. Asimismo, Wind y Engelhard (2012) incluyen mapas de variables que ilustran cómo la información orientada al contenido puede representarse visualmente para apoyar el sentido y el uso planificados de las puntuaciones de una evaluación mediada por observadores. Los mapas de variables son herramientas visuales importantes que pueden ser analizadas por ambas comunidades para ayudar a desarrollar un vocabulario común y una comprensión compartida sobre la evaluación de la escritura.

Otra fuente de evidencia sobre validez basada en la estructura interna de la evaluación incluye examinar el modo de evaluación empleado. Si la escritura se define en términos de ítems de respuesta seleccionada (por ejemplo, ítems de selección múltiple), existe una gran variedad de métodos psicométricos bien establecidos que pueden emplearse. Sin embargo, el uso de ítems de respuesta construida como los ensayos y los portafolios (evaluaciones que tienen más apoyo en la comunidad de investigación en escritura) requiere de datos adicionales sobre los evaluadores y el proceso de evaluación. La teoría de medición Rasch también entrega un conjunto de criterios claros que pueden usarse para evaluar la calidad de las evaluaciones (Engelhard, 2002). Como es bien sabido, otros modelos dentro de la teoría de respuesta al ítem no entregan un mapa simultáneo de las personas, los ítems y los evaluadores dentro de un único continuo subyacente (Engelhard, 2013).

La evidencia con respecto a relaciones con constructos conceptualmente asociados puede recogerse de variadas formas. La evidencia dentro de esta categoría incluye estudios que emplean las puntuaciones obtenidas mediante una evaluación de escritura como variables dentro de un marco más amplio, como una red nomotética (Whitely, 1983). Por ejemplo, Behizadeh y Engelhard (2014) desarrollaron un instrumento para examinar las percepciones de los estudiantes sobre la autenticidad de la escritura que podría estar relacionado con las puntuaciones en una evaluación de escritura. Además, el Funcionamiento Diferencial del Ítem [Differential Item Functioning (DIF)] (Engelhard, 2009) puede iluminar hasta cierto punto un impacto adverso que ocurra en la realidad (Kane, 2013) y que afecte a estudiantes de géneros, orígenes étnicos o niveles de logro específicos. De este modo, la teoría de medición Rasch ofrece herramientas analíticas que pueden servir para examinar los usos planificados y reales relacionados con las consecuencias de la aplicación de tests. Consideramos que los datos cuantitativos y visuales generados a partir de análisis Rasch podrían analizarse simultáneamente con datos cualitativos, como los derivados de entrevistas y observaciones obtenidas en estudios empíricos en salas de clases.

Las consecuencias de los tests basadas en evidencia pueden observarse de distintas maneras. Slomp et al. (2014) entregan un marco que debiera considerarse cuidadosamente como un prototipo para realizar estudios de validez consecuencial. La comunidad de investigación en medición estudia la validez consecuencial analizando el Funcionamiento Diferencial del Ítem y el Funcionamiento Diferencial de la Persona [Differential Person Functioning (DPF)]. Estos dos enfoques pueden usarse para examinar lo justo de las evaluaciones, incluyendo las consecuencias reales para los individuos y los subgrupos de estudiantes. Por ejemplo, Engelhard (2009) ha utilizado la teoría de medición Rasch para describir cómo el DIF y el DPF pueden emplearse para examinar y mejorar la validez de las interpretaciones de las puntuaciones para individuos y subgrupos de estudiantes.

En conclusión, creemos que la comunidad de investigación en medición puede ofrecer excelentes herramientas, como un plan de validación y mapas de variables, para recoger y analizar la amplia gama de datos necesarios para validar las evaluaciones de escritura. Sin embargo, nos parece que este proceso de validación estará incompleto a menos que la comunidad de investigación en escritura, así como los profesores, estudiantes y otras partes interesadas, sean parte de la empresa conjunta de desarrollar un sistema de evaluación válido y esta colaboración involucre la contribución de datos cualitativos importantes con respecto al impacto de las evaluaciones de escritura a gran escala sobre la enseñanza de la lectoescritura en las aulas.

Discusión y conclusiones

Los validadores también son una comunidad. Esto permite que sus miembros se repartan las tareas investigativas y educacionales de acuerdo a sus talentos, motivos e ideales políticos. La validación crecerá en proporción mientras hagamos nuestro mayor esfuerzo colectivo con nuestras mentes y corazones (Cronbach, 1988, p. 14).

El presente estudio contribuye a clarificar el concepto de validez e incluye sugerencias sobre futuras colaboraciones entre dos comunidades (medición y escritura), que se enfocan en asuntos relacionados con la evaluación de la escritura. Consideramos las siguientes preguntas:

1. ¿En qué consiste una evaluación válida de la escritura según la comunidad de investigación en escritura?
2. ¿En qué consiste una evaluación válida de la escritura según la comunidad de investigación en medición?
3. ¿Cuáles son algunos puntos de consenso y desacuerdo sobre el concepto de validez en ambas comunidades?

La comunidad de investigación en escritura respondería la primera pregunta enfocándose en la validez de constructo y consecuencial. Específicamente, dicha comunidad (a) articularía una visión sociocultural de la escritura que concibe el constructo de la escritura como un conjunto de prácticas dependientes del contexto y (b) destacaría la importancia de la evidencia sobre validez consecuencial derivada de la investigación de consecuencias reales, como la fuente más relevante de evidencia sobre validez. La comunidad de investigación en medición respondería la segunda pregunta refiriéndose a la definición consensuada de validez incluida en los *Estándares para Tests.* Algunos miembros de la comunidad de investigación en medición, como se demuestra en el presente artículo, conciben el proceso de validación de un modo que difiere de la definición consensuada. El enfoque basado en argumentos de Kane (2013) es congruente con la definición consensuada, mientras que Borsboom et al. (2004) se centran en un enfoque basado en atributos que se sostiene en examinar los efectos de los atributos, las variables latentes y los constructos sobre las variaciones en las respuestas de las personas.

La respuesta a la tercera pregunta tiene más matices. Existe consenso con respecto a la importancia de la validez de contenido y consecuencial, pero cada comunidad se especializa en distintas áreas que contribuyen al proceso de validación. Cuando conceptualizamos este estudio, esperábamos un mayor debate sobre validez entre las dos comunidades; en lugar de ello, descubrimos un nivel moderado de consenso con el potencial de generar una división de tareas más sistemática con respecto a cómo recoger y evaluar la validez de las evaluaciones de escritura. Sin embargo, una discusión crítica en la cual las comunidades de investigación necesitan involucrarse debe tener en cuenta cómo definir el constructo de la escritura. La comunidad de investigación en escritura consistentemente se basa en una comprensión sociocultural de la escritura, y su crítica a las evaluaciones estandarizadas de escritura es que estas no se alinean con dicha visión. Asimismo, otra discusión crítica para ambas comunidades se refiere al propósito de la evaluación de la escritura para estudiantes primarios y secundarios. Como se delinea en el presente artículo, la investigación en escritura hace énfasis en mejorar la enseñanza y el aprendizaje, mientras que la comunidad de investigación en medición ha recibido frecuentes peticiones de ayuda para desarrollar evaluaciones de nivelación y sumativas. Creemos que, al examinar múltiples herramientas ofrecidas por ambas comunidades, incluyendo el marco de validez consecuencial de Slomp et al. (2014) y el marco de validación de Sireci (2013), podemos colaborar para establecer tanto una definición consensuada de escritura como los propósitos de la evaluación de la escritura.

Finalmente, el presente artículo deja sin explorar los problemas de validación emergentes relacionados con la puntuación automática de ensayos [Automated Essay Scoring (AES)], como la investigación de Landauer, McNamara, Dennis y Kintsch (2013) y Shermis y Burstein (2003) y las críticas al AES, incluyendo Deane (2013) y Perelman (2014). Futuras investigaciones sobre la validez de la automatización de las prácticas de puntuación pueden usar como punto de partida las sugerencias para un marco integrado, representadas en la Tabla 1, de modo tal de involucrar a todas las partes interesadas en el proceso de validación.

Hay mucho que aprender de la *experticia* disciplinaria de cada comunidad abocada a la evaluación de la escritura. Cada comunidad tiene sus propias características para ofrecer, y el campo de la evaluación de la escritura se beneficiará de lo que puedan contribuir las comunidades de investigación en escritura y medición. Esperamos con ansias la nueva iteración epistemológica de la definición del concepto de validez, que cuente con la participación de voces fuertes de estas comunidades de práctica separadas, pero a la vez coincidentes.

El artículo original fue recibido el 19 de diciembre de 2014
El artículo revisado fue recibido el 14 de abril de 2015
El artículo fue aceptado el 13 de mayo de 2015

Referencias

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Bauer, E. B., & Garcia, G. E. (2002). Lessons from a classroom teacher's use of alternative literacy assessment. *Research in the Teaching of English 36*(4), 462-494.

Behizadeh, N. (2014). Mitigating the dangers of a single story: Creating large-scale writing assessments aligned with sociocultural theory. *Educational Researcher*, *43*(3), 125-136. doi: 10.3102/0013189X14529604

Behizadeh, N., & Engelhard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment. *Assessing Writing, 16*(3), 189-211. doi: 10.1016/j.asw.2011.03.001

Behizadeh, N., & Engelhard, G. (2014). Development and validation of a scale to measure perceived authenticity in writing. *Assessing Writing, 21*, 18-36. doi:10.1016/j.asw.2014.02.001

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071. doi: 10.1037/0033-295x.111.4.1061

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement*, *50*(1), 110-114. doi: 10.1111/jedm.12006

Broad, B. (2000). Pulling your hair out: Crises of standardization in communal writing assessment. *Research in the Teaching of English, 35*(2), 213-260.

Center for the Study of Testing, Evaluation, & Educational Policy (1992). *The influence of testing math and science in grades 4-12* (Vols. 1-5). Boston: Autor.

Cronbach, L. J. (1971). Test validation. En R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. En H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cronbach, L. J. (1990). *Essentials of psychological testing.* (5th ed.). Nueva York, NY: Harper.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7-24. doi: 10.1016/j.asw.2012.10.002

Duckor, B. M., Draney, K., & Wilson, M. (2009). Measuring measuring: Toward a theory of proficiency with the constructing measures framework. *Journal of Applied Measurement, 10*(3), 296-319.

Elliot, N., Deess, P., Rudniy, A., & Joshi, K. (2012). Placement of students into first-year writing courses. *Research in the Teaching of English, 46*(3), 285-313.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349. doi: 10.1037/1040-3590.8.4.341

Engelhard, G. (2002). Monitoring raters in performance assessments. En G. Tindal, & T. Haladyna (Eds.), *Large-scale Assessment Programs for ALL Students: Development, implementation, and analysis* (pp. 261-287). Mahwah, NJ: Erlbaum.

Engelhard, G. (2009). Using IRT and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement, 69*(4), 585-602. doi: 10.1177/0013164408323240

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Nueva York: Routledge.

Engelhard, G., & Behizadeh, N. (2012). Epistemic iterations and consensus definitions of validity. *Measurement: Interdisciplinary Research and Perspectives*, *10*(1), 55-58. doi: 10.1080/15366367.2012.681974

Engelhard, G., & Wind, S. A. (2013). Educational testing and schooling: Unanticipated consequences of purposive social action. *Measurement: Interdisciplinary Research and Perspectives, 11*(1-2), 30-35. doi: 10.1080/15366367.2013.784156

Gulliksen, H. (1950). *Theory of mental tests*. Nueva York, NY: Wiley.

International Reading Association (IRA), & National Council of Teachers of English (NCTE) (2010). *Standards for the assessment of reading and writing*. (IRA Stock No. 776; NCTE Stock No. 46864). EE.UU.: Autor.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527-535. doi: 10.1037//0033-2909.112.3.527

Kane, M. T. (2006). Validation. En R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education and Praeger.

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

Kelley, T. (1927). *Interpretation of educational measurements*. Yonkers, NY: World Book Co.

Ketter, J., & Pool, J. (2001). Exploring the impact of a high-stakes direct writing assessment in two high school classrooms. *Research in the Teaching of English, 35*(5), 344-393.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2013). *Handbook of latent semantic analysis.* Nueva York, NY: Psychology Press.

Madaus, G. F. (1994). A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy. *Harvard Educational Review, 64*(1), 76-95. doi: 10.17763/haer.64.1.4q87663r0j76rwv1

Messick, S. (1989). Meaning and values in test validation. The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11. doi: 10.3102/0013189x018002005

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi: 10.1037//0003-066x.50.9.741

Moss, P. A. (1994). Can there be validity with reliability? *Educational Researcher, 23*(2), 229-258. doi: 10.3102/0013189x023002005

Murphy, S. (2007). Culture and consequence: The canaries in the coal mine. *Research in the Teaching of English, 42*(2), 228-244.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, *10*(1), 1-29. doi: 10.1080/15366367.2012.669666

Perelman, L. (2014). When the «state of the art» is counting words. *Assessing Writing, 21*, 104-111. doi: 10.1016/j.asw.2014.04.001

Perry, K. (2012). What is literacy? A critical overview of sociocultural perspectives. *Journal of Language and Literacy Education*, *8*(1), 50-71. Recuperado de http://jolle.coe.uga.edu/wp-content/uploads/2012/06/What-is-Literacy_KPerry.pdf

Poe, M. (2014). The consequences of writing assessment. *Research in the Teaching of English, 48*(3), 271-275.

Prior, P. (2006). A sociocultural theory of writing. En C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 54-66). Nueva York, NY: Guilford Press.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. (Edición expandida, Chicago: University of Chicago Press, 1980).

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A crossdisciplinary perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, *50*(1), 99-104. doi: 10.1111/jedm.12005

Slomp, D. H., Corrigan, J. A., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study. *Research in the Teaching of English*, 48 (3), 276-302.

Thorndike, E. L. (1914). The measurement of ability in reading. *Teachers College Record, 15*(4), 207-277.

Thorndike, E. L. (1919). *An introduction to the theory of mental and social measurements. Revised and enlarged edition*. Nueva York, NY: Teachers College, Columbia University.

Thurstone, L. L. (1931). *The reliability and validity of tests*. Ann Arbor, MI: Edwards.

Wenger, E. (1998). *Communities of practice: Learning, meaning and identity*. Cambridge, UK: Cambridge University Press.

Wenger, E. (2010). Communities of practice and social learning systems: The career of a concept. En C. Blackmore (Ed.), *Social learning systems and communities of practice* (pp. 179-198). Londres: Springer Verlag and the Open University.

Wenger, E. (2015). *Communities of practice: A brief introduction*. Recuperado de http://wenger-trayner.com/introduction-to-communities-of-practice/

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197. doi: 10.1037/0033-2909.93.1.179

Wilson, M. R. (2005). *Constructing measures: An item response theory approach*. Mahwah, NJ: Lawrence Erlbaum.
Wind, S. A., & Engelhard, G. (2012). Evaluating the quality of ratings in writing assessment: Rater agreement, precision, and accuracy. *Journal of Applied Measurement, 13*(4), 321-335.