Para citar este artículo, le recomendamos el siguiente formato:

Medina, L. (2013). La evaluación en el aula: reflexiones sobre sus propósitos, validez y confiabilidad. *Revista Electrónica de Investigación Educativa, 15*(2), 34-50. Recuperado de http://redie.uabc.mx/vol15no2/contenido-medina.html

Revista Electrónica de Investigación Educativa

Vol. 15, Núm. 2

La evaluación en el aula: reflexiones sobre sus propósitos, validez y confiabilidad[1]

Classroom-Based Assessment: Reflections on its Purposes, Validity and Reliability

Luis Medina Gual
luis.medina@semperaltius.org

Departamento de Educación, Universidad Iberoamericana

Av. Lomas Anáhuac 46
Col. Lomas Anáhuac C.P. 53786
Huixquilucan, Estado de México, México

(Recibido: 25 de abril de 2012; aceptado para su publicación: 21 de enero de 2013)

Resumen

El presente trabajo tiene como fin presentar un análisis sobre la evaluación en el aula a través de la discusión de sus propósitos, validez y confiabilidad. En este sentido, se delimitan los propósitos con base en los actores y conceptos de calidad, para posteriormente plantear los criterios de validez y confiabilidad a la luz de la naturaleza del proceso educativo. El artículo propone tres tipos de estrategias para la valoración de la validez y confiabilidad de la evaluación en el aula.

Palabras clave: Evaluación del aprendizaje, evaluación del estudiante, validez.

[1] Las traducciones literales de los textos en inglés fueron realizadas por el autor del artículo.

Abstract

This paper presents an analysis of classroom-based assessment by means of a discussion of it s purposes, validity and reliability. To this end, its purposes were defined based on the stakehol ders and concepts of quality; subsequently, validity and reliability criteria were posited in light of the nature of the educative process. The article proposes three types of strategies for appraising the validity and reliability of classroom-based assessments.

Keywords: Learning assessment, student evaluation, validity.

I. Introducción

La relación entre la psicología y la educación ha sido y sigue siendo enriquecedora y conflictiva al mismo tiempo, "sobre todo porque a la hora de alimentarse recíprocamente, no siempre se han tenido las precauciones necesarias para evitar las transpolaciones mecánicas" (Bixio, 2006, pp. 119-120).

El caso de la evaluación del aprendizaje no es la excepción. Desde hace casi 70 años, Scates (1943) reflexionó sobre las diferencias y similitudes que existían entre lo que identificó como la medición científica y lo que hace un docente en el aula para evaluar a sus alumnos:

> (…) los científicos están interesados en la verdad para lograr generalizaciones, mientras que los profesores buscan información por su valor práctico; (…) los especialistas en medición no pueden medir de forma constante [a los sujetos] pero el profesor necesita y debe hacerlo; el científico mide rasgos a través de una escala pero el profesor mide el desarrollo [o aprendizaje] en etapas; y el especialista de la medición, mide habilidades formales [en situaciones controladas o asépticas] a través de pruebas, pero al docente le importan las dinámicas y conductas de sus estudiantes en situaciones diarias.

Históricamente el desarrollo en la psicometría se ha realizado en contextos de las pruebas a gran escala, por expertos en estadística y rara vez por docentes o expertos de una disciplina (Shepard, 2001), y posteriormente se han "transferido" dichos conocimientos al contexto de la evaluación del salón de clases (Brookhart, 2003). En otras palabras, en lugar de desarrollar una teoría y pensamiento propios y emergidos del salón de clase, se procedió a "pedir prestada" la teoría formulada para contextos de la evaluación de gran escala o masiva (Brookhart, 2003). Ello se evidencia en el hecho de que los estándares desarrollados históricamente por la American Psychological Association (APA), la American Educational Research Association (AERA) y el National Council on Measurement for Education (NCME) desde 1966, no se ocupan de las evaluaciones emanadas del aula o elaboradas por docentes de forma directa (Cronbach, 1998), y no fue sino hasta el año 2003 que el Joint Commitee on Standards for Educational Evaluation elaboró un documento que explícitamente atendía a las evaluaciones en el aula; el mismo documento indica que éste es el primer intento de discusión sobre cómo planear, conducir, usar y valorar las evaluaciones en el aula (Joint Commitee, 2003).

Si bien, tanto la evaluación en contextos escolares como la estandarizada tienen como función básica el determinar cuánto sabe una persona o qué tan bien puede poner en práctica determinadas habilidades (Aiken, 1996), la literatura señala dos áreas de tensión:

Sobre la naturaleza del acto educativo: Donde (1) la mayoría de las teorías del test trabajan con muestras de gran tamaño mientras que los docentes tienen un tamaño proporcionalmente pequeño de estudiantes (Brookhart, 2003), (2) la evaluación a gran escala evalúa contenidos o habilidades genéricos o comunes (pensamiento o comprensión) mientras que la evaluación escolar tiene gran interés en contenidos curriculares disciplinares o muy específicos y propios de un curso y una metodología didáctica (Aiken, 1996), (3) la evaluación en el salón de clases y la instrucción se encuentran estrechamente ligadas (Brookhart, 2003; McMillan, 2003) por lo que (4) la evaluación en el aula es esencialmente de carácter formativo (Brookhart, 2003; McMillan, 2003).

Diferencias sobre las nuevas corrientes psicopedagógicas: Autores como Shepard (2000, p. 1) afirman que: "La disonancia actual (…) emerge por la incompatibilidad entre las viejas ideas de los tests y las ideas transformadas de la enseñanza", y ello adquiere especial relevancia si se considera la premisa de De la Orden (2011) que sugiere que el enfoque de la educación determina el qué evaluar (su objeto), el cómo evaluar (su método) y el cuándo evaluar (su planeación). En otras palabras, "lo que evalúas es lo que obtienes" (Shepard, 2000, p. 36).

Al respecto resulta de interés que como Ayers (1918 en Shepard, 2000) argumenta, no es ninguna coincidencia que Edward Thorndike fuera tanto el creador de la teoría asociacionista del aprendizaje como el "padre" de la "medición científica". Como muestra de lo anterior es posible advertir cómo las teorías del test parecen acercarse en mayor medida a teorías asociacionistas ("Greeno, Collins y Resnick, 1996; Shepard, 1996; Shepard, 1991b; Shulman y Quinlan (como se citó en Shepard, 2000).

Desde nuevas corrientes pedagógicas que apoyan a movimientos como la instrucción diferenciada (Tomlinson e Imbeau, 2010), el "Understanding by design" (Shepard, 2000; Wiggins y McTighe, 2005), las corrientes de "auto, co y hetero evaluación" que hacen responsable al alumno de la evaluación misma (Hidalgo, 2005) o el "aprendizaje para todos" y el "aprendizaje para toda la vida" (OCDE, 2012) se pueden pensar como opuestas a algunas nociones compartidas por casi la totalidad de las teorías de la medida (como la "discriminación" y "dificultad" de un reactivo) debido a que en estas corrientes, el propósito de la evaluación yace en que el alumno demuestre lo que ha aprendido a través de los medios que le den más ventaja (Shepard, 2000). Esto contrasta con la idea de que la evaluación tiene que ser idéntica para todos en aras de la "justicia" y la "objetividad" de la medición (McMillan, 2003; Shepard, 2000).

Ante esto la literatura revela tres caminos: desarrollar el razonamiento probabilístico a situaciones de evaluación para el salón de clases (Mislevy, 2004), adaptar las teorías de la medición al aula (Brookhart, 2003; Li, 2003), o desarrollar una teoría alternativa que emerja desde la práctica docente en el salón (Moss, 1994 y 2004), forjando una

teoría única que pertenezca al aula (Brookhart, 2003). Incluso se advierte la posibilidad de combinar las tres posturas (Mislevy, 2004; Shepard, 2000).

Lo cierto es que la discusión continúa y aunque algunas veces es acalorada (Li, 2003; Mislevy, 2004; Moss, 1994; Moss, 2004) tiene como único propósito proporcionar a los docentes herramientas y estrategias para desarrollar "evaluaciones de calidad" (Stiggins, 2001 en Brookhart, 2003). En coherencia con lo anterior, el presente trabajo tiene por objeto el reflexionar sobre la idea de la calidad de la evaluación en el aula a través de la discusión de sus propósitos, validez y confiabilidad.

II. El propósito de la evaluación en el aula

Existe un relativo consenso en plantear que el mejor punto de partida para definir la "evaluación del aprendizaje de calidad" es la definición de los propósitos sustantivos de este proceso (Brookhart, 2009, 2011; De la Orden, 2011). Con este fin, Brookhart (2011, p. 12) propone reflexionar sobre las preguntas: "¿Qué significados queremos que la evaluación tenga?" y "¿Quién(es) es/son la(s) audiencia(s) primaria(s) de los mensajes que se derivan de la evaluación?" y ello complementa a la función que señala De la Orden (2009, en De la Orden 2011, p. 2) para la evaluación en la educación: "optimizar su estructura [(de la educación)], proceso y producto actuando como un mecanismo de retroalimentación de tales sistemas para asegurar su permanencia, eficacia y funcionalidad".

Por lo tanto, la evaluación del aprendizaje debería servir al propósito de ser un mecanismo que, desde una perspectiva sistémica, vigile las diferentes "relaciones de coherencia" (De la Orden *et al.*, 1997) entre insumos, procesos, productos, contextos y propósitos/metas en aula, a través de la dotación de "significados" pertinentes y dirigidos –a los diferentes actores implicados– en pos de su actuación para la mejora continua. Dichas "relaciones de coherencia" (De la Orden *et al.*, 1997) se traducen en concepciones de calidad como: eficacia, eficiencia y pertinencia-relevancia (Muñoz-Izquierdo, 2009).

En este sentido, el siguiente cuadro tiene como propósito clarificar la premisa anterior al clasificar diferentes "propósitos" de la evaluación del aprendizaje con base en las "relaciones de coherencia" de De la Orden *et al.* (1997; 2011) al conceptualizarlos como "tipos de calidad" (Muñoz-Izquierdo, 2009) y de los actores que deberían de ser considerados (Brookhart, 2011):

Tabla I. Propósitos de la evaluación del aprendizaje en el aula

Actor	Eficacia	Eficiencia	Pertinencia-relevancia
Alumno	Satisfacer necesidades de información sobre su rendimiento en términos de "mejora en su desempeño" (Stiggins, 2004) para la mejora de sus estrategias de aprendizaje en próximas ocasiones (Chappuis y Stiggins, 2002)	Conocer: qué pretende el docente y el currículum, qué tanto debe trabajar, qué debe reforzar (Stiggins, 1992) para lograr manejar y mejorar el proceso de aprendizaje (Chappuis y Stiggins, 2002)	Determinar la cercanía de la instrucción y los contenidos curriculares con las necesidades de su comunidad y con las demandas del mundo actual
	Determinar el logro de la adquisición y de manejo de "códigos académicos" (Bernstein en Sadovnick, 2001) y valorar sus consecuencias en el corto y largo plazo		
Docente	Corroborar el logro de aprendizajes en los alumnos (Brookhart, 2009) para la mejora continua de su labor docente (Chuappuis y Stiggins, 2002)	Corroborar la didáctica e instrucción (Brookhart, 2009) y tomar decisiones en el aula durante el período de clases (Brookhart, 2004; Stiggins, 1992)	Adaptar la instrucción y currículum según necesidades de los estudiantes y la sociedad
Padres de familia	Contrastar las expectativas del desempeño de sus hijos con el desempeño reportado (Stiggins, 1992) y orientarlos para su futuro	Dar seguimiento y apoyo continuo al desempeño de sus hijos	Valorar los aprendizajes de sus hijos para orientarlos en un oficio o profesión coherentes con las necesidades de su comunidad y con las demandas del mundo actual
	Determinar el logro de la adquisición y de manejo de "códigos académicos" (Bernstein en Sadovnick, 2001) y valorar sus consecuencias en el corto y largo plazo		
Autoridades escolares (internas)	Corroborar los logros del aprendizaje en el aula	Corroborar y dar seguimiento al proceso de enseñanza-aprendizaje de una asignatura el particular	Valorar el currículum en tanto su pertinencia-relevancia para su comunidad y/o para estudiantes en específico
Actores externos	Corroborar el logro de aprendizajes para la selección de estudiantes (Ej. Universidad)	Corroborar y dar seguimiento al proceso de enseñanza-aprendizaje en una institución educativa	Valorar el currículum en tanto su pertinencia-relevancia para la comunidad

Una vez que se clarifica el tipo de finalidades que se desea, se deberá replantear el significado de los procesos mismos de evaluación (Brookhart, 2003). Ello remite a la existencia de "características", "criterios" o "elementos" que el proceso de evaluación, sus herramientas, instrumentos y estrategias deberían de poseer. Algunos de éstos, que emergen de la literatura sobre la evaluación del aprendizaje en el aula son: "precisión", "especificidad", "momento adecuado de aplicación" (Reeves, 2011), "justicia" (Brookhart, 2004; Camara y Lane, 2006; Reeves, 2011) y "utilidad" (Brookhart, 2004). Sin embargo, se continúa considerando como imprescindibles para la evaluación en aula a la "validez" y "confiabilidad" (Allen, 2005; Brookhart, 2004; De la

Orden, 2011; Gullickson, 2003; Hidalgo, 2005; Hopkins, 1998; Li, 2003; McMillan, 2003; Mislevy, 2004; Moss, 1994; Moss, 2004; Shepard, 2000; Stiggins, 2004; Teasdale y Leung, 2000). Es por esto que se ha optado por que las siguientes secciones del presente trabajo tengan como cometido el repensar estos conceptos en el aula.

III. Validez en el aula

Históricamente, la "validez" ha sufrido una metamorfosis o evolución durante la segunda mitad del siglo XXI (Goodwin y Leech, 2003). En un primer momento, se conceptualizaba en función a la capacidad de correlación de los puntajes de un test con algún criterio externo (Guilford, 1946 en NCME-ACE, 1993). Con el surgimiento de los primeros estándares elaborados por la APA-AERA-NCME (1966), la validez fue entendida en función al grado en que el test produce información útil y con un propósito específico. Es en este primer documento de los estándares, que se hace alusión a la idea de tres tipos de validez (contenido, criterio y constructo) propuesta por Cronbach y Meehl en 1955 (Goodwin y Leech, 2003). Posteriormente, fue el mismo Cronbach quien en 1980 comenzó a desdibujar la necesidad de "unificar" a la validez, argumentando que la esencia de ésta se encontraba en la validez de constructo (NCME-ACE, 1993). Ello condujo a que, en el año del 1985, los nuevos estándares de la APA-AERA-NCME advirtieran que el uso de tres categorías o etiquetas de la validez no debería conducir a inferir la existencia de tres tipos de validez (Goodwin y Leech, 2003). Finalmente fue en 1999 que la APA-AERA-NCME remplazó la idea de validez de contenido, criterio y constructo a una validez que puede ser valorada a través de cinco tipos de evidencias: del contenido del test, de los procesos de respuesta, de su estructura interna, de la relación con otras variables y de las consecuencias producto de su aplicación e interpretaciones (Goodwin y Leech, 2003). Así pues, el concepto de validez volvió a unificarse (Jonson y Plake, 1998; NCME-ACE, 1993) al entenderla como un criterio multidimensional y complejo (Goodwin y Leech, 2003) que se define como el grado en que la evidencia y teoría apoyan a la interpretación de los puntajes derivados de un test, sus usos y propósitos (APA-AERA-NCME, 1999).

A pesar de ello, al día de hoy el término de validez parece ser usado de forma distinta por diferentes autores, en diferentes contextos (Pedhazur y Pedhazur, 1991) y no necesariamente alineada (en la teoría o en la praxis) a los estándares (Jonson y Plake, 1998).

Ahora bien, como ya se ha comenzado a discutir, algunas consideraciones de la evaluación en el aula son que esta debe: (1) tener el significado de informar decisiones instruccionales –de carácter formativo– (Brookhart, 2005; McMillan, 2003); (2) no necesariamente aludir a inferencias sobre un proceso interno a un sujeto sino sobre un sujeto en relación con otros y que muchas veces es de suma importancia el entender la evaluación en un contexto (Brookhart, 2003; Moss, 2004); (3) no necesariamente tiene que ser el mismo contexto o ambiente para todos los evaluados (Moss, 2004).

Es por esto que diferentes autores sugieren que una evidencia de especial relevancia que se debería de retomar es la coherencia entre los criterios y modos de evaluación,

el proceso didáctico, el currículum y la finalidad (o significado) misma de la evaluación en un contexto determinado (Brookhart, 2003; Brookhart, 2004; De Caimillani *et al*., 2001; De la Orden, 2011; McMillan, 2003; Smith, 2003). En otras palabras y como afirma De la Orden (2011), la validez de la información, interpretaciones y consecuencias de la evaluación del aprendizaje en el salón de clase implicaría una validez "axiológica", "curricular" e "instrumental". Por ello, la validez en el aula no podría ser determinada de manera absoluta sino en relación con su adecuación a los propósitos y situación específica de aplicación (De Camillan *et al.*, 2001; Moss, 2004; Winter, 2000) y de los contextos en que los alumnos son evaluados y que, por tanto, influyeron en los desempeños (Moss, 2004). Así pues, la validez en contextos escolares debería hacer un especial énfasis en los análisis e interpretaciones hechas sobre el rendimiento de un estudiante (Hidalgo, 2005; Joint Commitee, 2003; Moss, 2004).

Y, si la validez se encuentra en relación estrecha con el coadyuvar al proceso educativo, preocupaciones como "cuántas evaluaciones son necesarias para la representatividad" pasan a un segundo término en tanto que la interacción del contexto del salón y el currículum aseguran la validez de lo que es evaluado (Teasdale y Leung, 2000).

Por tanto, se plantearía valorar la validez de la evaluación en el aula a través de estrategias interpretativas para contextos sociales particulares donde no se considere como deseable o posible la estandarización (Moss, 2004), sino que más bien, los métodos de evaluación sean relevantes y representativos a los contenidos evaluados (Joint Commitee, 2003). Para lograr lo anterior, una alternativa es la "hermenéutica" como estrategia interpretativa de la información. Este enfoque se vería complementado a través del uso de estrategias para el rigor en investigaciones cualitativas (Ali y Yusof, 2011; Barusch, Gringer y George, 2011; Denzin y Lincon, 2005; Golafshani, 2003), o a través de estrategias propias del ambiente del salón de clases como entrevistas a los alumnos y padres de familia (Moss, 2004).

Por otra parte, también se propone valorar la validez a través de "las consecuencias de la evaluación" (Moss, 2004). En otras palabras, se le plantearía como el logro de consecuencias deseadas para el aprendizaje, la instrucción y sociales, donde a mayor importancia tenga una evidencia para la toma de decisiones, se debería vigilar con mayor rigor su validez (Moss, 2004; Joint Commitee, 2003).

Por todo lo discutido, si la validez en el aula se conceptualiza como aquel criterio con el que se verifica que las evaluaciones están desarrolladas e implementadas de forma tal, que las interpretaciones sobre los estudiantes no estén abierta a malinterpretaciones (Joint Commitee, 2003, p. 127), se podría pensar en la existencia de tres tipos de evidencia para su valoración:

Evidencias curriculares o de concreción curricular, que valorarían el grado en que los datos o mediciones reflejen lo que se pretende conocer del estudiante (Álvarez-Gayou, 2010; Brookhart, 2009; Cronbach, 1980 en NCME-ACE, 1993; Hernández, Collado y Bapstista, 2007; Denzin y Lincon, 2005; Winter, 2000).

Evidencias interpretativas, que tendrían a bien el valorar que las interpretaciones o inferencias a partir de los datos reflejen lo que se pretendía conocer (APA-AERA-NCME, 1985; APA-AERA-NCME, 1999; Brookhart, 2003; Cronbach, 1998; Denzin y Lincon, 2005; Hopkins, 1998; NCME-ACE, 1993; Moss, 2004) y que tiene como consecuencia también el sopesar la importancia de la rigurosidad de la interpretación (Denzin y Lincon, 2005; Miles y Huberman, 1984).

Evidencias instrumentales, que valorarían la existencia de consecuencias didácticas, instruccionales y/o sociales deseables. Este tipo de evidencias se retomada autores clásicos como Cronbach (“1980, 1988” como se citó en Moss, 1994;), Thorndike y Hagen (1996), Linn *et al.* (1993) y los estándares de la APA-AERA-NCME (1996; 1985, 1999). También se sugiere la importancia de que la validez demuestre cómo cumple una función para la que fue pretendida inicialmente (Eisner, 1998; Gronlund, 1971; Hopkins, 1998).

Al igual que en el caso de los estándares de APA-AERA-NCME (1999) definen diferentes estrategias para valorar la validez en las puntuaciones e interpretaciones emanadas de pruebas de gran escala, la tabla II tiene como cometido mostrar las evidencias y estrategias para su valoración en el aula:

Tabla II. Evidencias y estrategias para la valoración de la validez en evaluaciones en el aula

Evidencia para la valoración de la validez	**Definición de la evidencia**	**Estrategias para el docente o tutor**
Evidencias curriculares (De la Orden, 2011; De Camillan *et al*., 2001)	Evidencias que permiten valorar el grado de coherencia entre la evaluación propuesta y el currículum prescito y real.	Buscar la representatividad del contenido evaluado (Li, 2003); lograr que las calificaciones contemplen únicamente desempeños de corte académicos y no evalúen esfuerzo o interés (Allen, 2005); justificar la coherencia entre el tipo de instrumentos utilizados y la didáctica del docente.
Evidencias interpretativas	Evidencias sobre sobre las interpretaciones e inferencias derivadas de la evaluación; deberían corroborar el rigor de las estrategias interpretativas empleadas según el contexto social y el actor implicado (Brookhart, 2003; Moss, 2004).	Uso de la hermenéutica como estrategia interpretativa (García, 2002; Hidalgo, 2005; Moss, 2004) y de estrategias que apuntalen el rigor metodológico de la interpretación en la evaluación (Ali y Yusof, 2011; Barusch, Gringer y George, 2011; Golafshani, 2003; Moss, 2004; Quin, 1990; Denzin y Lincon, 2005); uso de varios instrumentos y metodologías de evaluación (Brookhart, 2009; Joint Commitee, 2003); determinación de posibles influencias sistemáticas (bias) en las evaluaciones (Gronlund, 1971; Joint Comitee, 2003); lograr que las calificaciones contemplen únicamente desempeños de corte académicos y no evalúen esfuerzo o interés (Allen, 2005); preparar una guía para la calificación y describir y justificar el proceso de interpretación (Joint Commitee, 2003).
Evidencias instrumentales	Evidencias que favorecen la reducción de la diferencia entre el desempeño actual del estudiante y las metas educativas planteadas (Chappuis y Stiggins, 2002) o sociales, al maximizar las consecuencias positivas de la evaluación y minimizar las negativas (Brookhart, 2004)	Que se promueva las consecuencias deseadas para el aprendizaje y la instrucción considerando como criterio de validez de una evidencia de aprendizaje su relación con la toma de decisiones (Moss, 2004); que exista una correlación entre los resultado de las evaluaciones y un evento futuro (Barbara y leydens, 2000; Gronlund, 1971), que el tipo de instrumento y su interpretación sea coherente con los propósitos de la evaluación (Joint Commitee, 2003).

IV. La confiabilidad en el aula

A la confiabilidad se le reconoce un papel de especial importancia social y científica (Parkes, 2007). De igual forma que la validez, varios académicos solicitan la clarificación del concepto de confiabilidad como una propiedad de un conjunto de puntajes o interpretaciones y no así de los instrumentos o tests (Brennan, 2011; Fribie, 2005).

De forma general es posible advertir que la confiabilidad es entendida como “la estabilidad o consistencia de los resultados o interpretaciones” (Gronlund, 1971; NCME-ACE, 1993; Salkind, 2006; Vogt, 1993) respondiendo con ello a las preguntas: “si la

gente fuera examinada dos veces, ¿Coincidirían las dos puntuaciones? ¿Hasta qué punto?" (Cronbach, 1998, p. 219). Dichos resultados deberían de ser independientes a la influencia de factores que podrían alterar la medición (Brookhart, 2003). Desde esta interpretación surge el concepto de error de la medida (sistemático o asistemático) que debería de ser minimizado o evitado (Pedhazur y Pedhazur, 1991; Vogt, 1993) con el fin de vigilar la exactitud y precisión del procedimiento de medición (Thorndike y Hagen, 1996). Muchas veces, la confiabilidad se piensa como equivalente a los índices que lo representan (Fernández, Noelia y Pérez, 2009).

Ahora bien, específicamente para las evaluaciones del aula, valdría la pena preguntarse: "¿Qué significaría entonces la confiabilidad de una medición o interpretación que sólo se va a administrar una única vez para una única persona?" (Smith, 2003). En este sentido, resulta interesante la definición de la confiabilidad de la evaluación en el aula propuesta por el Joint Commitee (2003): el grado de consistencia entre las puntuaciones o información obtenida de un proceso de recolección de datos.

Coherente con lo anterior, algunos autores argumentan que las evaluaciones estandarizadas –y externas– deben de ser más confiables en tanto que sirven para tomar decisiones más importantes (para todo un sistema educativo), mientras que la confiabilidad de la evaluación del aprendizaje en el aula no es de suma relevancia al contexto escolar debido a que es posible corregir malas decisiones del docente en el día a día o minuto a minuto, gracias a la nueva información que emerge constantemente del mismo salón (Brookhart, 2003; Joint Commitee, 2003; Shepard, 2000; Smith, 2003). Sin embargo no se desdeña este criterio debido al reconocimiento de que una mala señal para el alumno serían el ofrecerle datos o interpretaciones erráticas por parte del docente (Shepard, 2000).

A pesar de lo anterior, las teorías de la medida de pruebas a gran escala tienen como supuestos la independencia de las respuestas, en su mayoría no contemplan al evaluador como parte del cálculo de la confiabilidad (Moss, 1994) y si se pensara en la confiabilidad como el logro de consistencia entre dos o más mediciones, este criterio pareciera ser la antítesis misma del propósito central del acto educativo: el aprendizaje (Lamprinou y Christie, 2009; Parkes, 2007).

Otros ejemplos de la falta de coherencia entre las teorías del test y las evaluaciones en el aula son el hecho de que: (1) en la teoría clásica, la diferencia entre los puntajes de una aplicación y otra no sería consideradas como aprendizaje sino como un error en la administración o un efecto de la práctica sobre el test (Lamprinou y Christie, 2009; Smith, 2009); (2) en las teorías de la medida los reactivos deberían poder discriminar entre los sujetos que saben y no saben (Tristán, 1998; Chávez y Saade, 2010). Es decir, un ítem que todos los estudiantes respondan bien o mal, no contribuiría a la confiabilidad (Smith, 2003). En este sentido, este tipo de ítems (que por ejemplo, obtendrían correlaciones biseriales cercanas a 0.0) deberían de ser descartados bajo las sugerencias clásicas de la teoría del test (cfr. Chávez y Saade, 2010). Una vez más, esto iría en contra de lograr el "aprendizaje para todos" o de movimientos instruccionales como el "Mastery Learning" (Araisan, Bloom y Carroll, 1971). Por ello, habría que cuestionarse si, al menos en contextos áulicos, existiría la posibilidad de

definir (o incluso fijar) los parámetros o la probabilidad de ejecución de un ítem o de un sujeto. Y, por último, (3) tanto la teoría clásica como la de generazibilidad y la de respuesta al ítem funcionan bajo el supuesto de unidimensionalidad, donde los constructos de interés son ortogonales (no se relacionan) (Parkes, 2007). En la educación y más aún con nuevas corrientes como la educación basada en competencias, esto implicaría ir en contra de la integración de los aprendizajes.

Sin embargo, existen ejemplos como en el caso de Marzano (2000), quien ha buscado repensar la confiabilidad y las calificaciones en el aula a través de propuestas como el cálculo del desempeño del estudiante por medio de funciones de potencia, de relativa facilidad de cómputo para el docente y que se asemejan con cómo la psicología educativa indica que aprende el ser humano. Otros ejemplos del mismo autor (Marzano, 2000) son las consideraciones y adecuaciones de los supuestos y aportaciones de corrientes como la teoría de respuesta al ítem para enriquecer los procesos de evaluación alineados a estándares. Estos esfuerzos no parecieran de ningún modo ser aislados y han continuado hasta el día de hoy, repensando la confiabilidad de la evaluación en el aula (Reeves, 2011). A pesar de esto y como señala Mislevy (2004), los retos de los especialistas del test para hacer “sentido” de la evidencia continúan.

Coherente con las alternativas anteriores y al igual que en el caso de la validez, sería posible de sugerir algunas evidencias y estrategias para la valoración de la confiabilidad de la evaluación en el aula:

Tabla III. Evidencias y estrategias para la valoración de la confiabilidad en evaluaciones en el aula

Evidencia para la valoración de la confiabilidad	**Definición de la evidencia**	**Estrategias para el docente o tutor**
Evidencias interpretativas	Grado de confianza que se tiene en que la calificación o interpretación de un estudiante refleja su desempeño real y es precisa (Brookhart, 2004).	Obtención de calificaciones o interpretaciones similares durante un periodo de tiempo (Gronlund, 1971); criterios de evaluación claros, uso de rúbricas analíticas (Brookhart, 2004; Barbara y Leydens, 2000; Moss, 1994; Stiggins, 1992); retomar evaluaciones de corte criterial (De la Orden, 2011; Smith, 2003); uso de ejemplos de trabajos para diferentes niveles de desempeño (Brookhart, 2004); incrementar las especificaciones para calificar (Moss, 1994; Reeves, 2011); consideración del contexto y factores que inciden en la evaluación y cómo influye en el desempeño (Ej. Vocabulario o redacción) (Barbara y Leydens, 2000; Gullickson, 2003); consideración del efecto halo (Gullickson, 2003; Shepard, 2000); descripción específica de los procedimientos de evaluación (Gullickson, 2003); entrevistas con padres de familia y otros actores (stakeholders) que valoren la interpretación (Joint Comitee, 2003); determinar el impacto de la evaluación y por tanto el tipo de instrumento que debiera ser utilizado (Joint Commitee, 2003); corroborar que el tipo de instrumento sea coherente con el tipo de interpretación deseada (Joint Commitee, 2003); evitar calificar cuando se está cansado o agotado (Joint Commitee, 2003); asumir que la confiabilidad garantiza la validez (Joint Comitee, 2003); uso de metodologías como la hermenéutica para la elaboración de interpretaciones (García, 2002; Moss, 1994).
Evidencias de estabilidad	Estabilidad de la información sobre la diferencia entre el desempeño real de un estudiante y el ideal o prescrito (Brookhart, 2003).	Todas las estrategias de incorporación de la fila anterior; coherencia entre la percepción del estudiante sobre las metas e instrucción solicitado por el docente y lo planeado y ejecutado por el profesor (Brookhart, 2003); intentar incrementar la longitud de la prueba, el número de calificadores o jueces, y el número de instrumentos y métodos de evaluación (Joint Comitee, 2003).
Evidencias instrumentales	Suficiencia de la información para la toma de decisioneas (Smith, 2003).	Todas las estrategias de incorporación de las filas anteriores; cuando un docente usa de forma consiente y defendible la información para apoyar el aprendizaje de sus alumnos y el proceso instruccional así como la toma de decisiones(Gullickson, 2003); representatividad de lo que se pretende evaluar (Brookhart, 2003; Smith, 2003); reflexión sobre las inconsistencias de datos o interpretaciones con base en evidencias (Parkes, 2007); decidir el grado de confianza según la importancia de la decisión (Parkes, 2007); asumir que la confiabilidad es la misma para todos los grupos y situaciones (Joint Comitee, 2003).

V. Conclusiones

El presente trabajo ha tenido como propósito central el reflexionar sobre la evaluación en el aula a través de la discusión de sus propósitos, validez y confiabilidad. Lo anterior, debido al hecho de que históricamente, la psicología ha aportado a la educación de teorías que no necesariamente surgieron en y para el aula. Una de estas áreas es la evaluación del aprendizaje.

En este sentido, una primera discusión es la delimitación de los propósitos sustantivos de la evaluación del aprendizaje en el aula que se definen con base en los diferentes actores del proceso educativo y de las concepciones de calidad y que condicionan o regulan la necesidad de su validez y confiabilidad.

Así, se procede a definir a la validez como un criterio de calidad complejo, que en el aula debe hacer especial énfasis en los análisis e interpretaciones hechas sobre el rendimiento de los estudiantes (Joint Commitee, 2003). En una misma tónica y coherente con las sugerencias de evidencias en lugar de tipos de validez propuestos en los estándares la APA-AERA-NCME (1999), se sugiere la existencia de tres tipos de evidencias que el docente o tutor prodría utilizar para valorar la validez de sus evaluaciones: curriculares, interpretativas e instrumentales.

Para finalizar, se discute el criterio de confiabilidad a la luz de las particularidades de evaluaciones en contextos áulicos. Como se podría pensar, debido al hecho de que las teorías del test no surgen para estos contextos, se evidencian algunas incoherencias con las necesidades propias de un salón de clases. A pesar de lo anterior, se muestran propuestas de autores americanos que han hecho adecuaciones a las mismas con el fin de vigilar su pertinencia para su uso en la cotidianidad escolar. Por tanto, se procede a hacer una propuesta de tres tipos de evidencias (interpretativas, de estabilidad e instrumentales) para valorar la confiabilidad de la evaluación en el aula.

A manera de cierre habría que mencionar que, a lo largo de este trabajo, se ha intentato realizar una reflexión sobre la calidad de la evaluación en el aula. Todo esto con el fin de evitar lo que en algunos campos de la educación se ha hecho: copiar o trasladar de forma mecánica, las teorías y propuestas de la psicología al contexto escolar.

Referencias

Ali, A. M. y Yusof, H. (2012). Quality in qualitative studies: The case of validity, reliability and generalizability. *Issues in Social and Environmental Accounting, 5*(1-2), 25-64.

Allen, J. D. (2005). Grades as valid measures of academic achievement of classroom learning. *The Clearing House, 78*(5), 218-223.

Álvarez-Gayou, J. L. (2010). *Cómo hacer investigación cualitativa: Fundamentos y metodología*. México: Paidós.

American Psychological Association, American Educational Research Association y National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association y National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Barusch, A., Gringeri, C. y George, M. (2011). Rigor in qualitative social work research: A review of strategies used in published articles. *Social Work Research, 35*(1), 11-19.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*(4), 295-317.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*(4), 5-12.

Brookhart, S. M. (2004). Assessment theory for college classrooms. *New Directions for Teaching & Learning*, *100*, 5-14.

Brookhart, S. M. (2009). The many meaning of "multiple measures". *Educational Leadership, 67*(3), 6-12.

Brookhart, S. M. (2011). Starting the conversation about grading. *Educational Leadership, 69*(3), 10-14.

Camara, W. J. y Lane, S. (2006). A historical perspective and current views on the standards for educational and psychological testing. *Educational Measurement: Issues & Practice, 25*(3), 35-41.

Chappuis, S. y Stiggins, R. J. (2002). Classroom assessment for learning. *Educational Leadership, 60*(1), 40.

Chávez, C. y Saade, A. (2010). *Procedimientos básicos para el análisis de reactivos.* México: Centro Nacional de Evaluación para la Educación Superior.

De Camilloni, A. R. W., Celman, S., Litwin, E. y Palou de Maté, M. C. (2001). *La evaluación de los aprendizajes en el debate didáctico contemporáneo.* Argentina: Paidós Educador.

De la Orden, A. (2011). Reflexiones en torno a las competencias como objeto de evaluación en el ámbito educativo reflections on competency based assessment in education. *Revista Electrónica de Investigación Educativa, 13*(2), 2.

De la Orden, A., Asensio, I., Carballo, R., Fernández Díaz, J., Fuentes, A., García Ramos, J. *et al.* (1997). Desarrollo y validación de un modelo de calidad universitaria como base para su evaluación. *Revista Electrónica de Investigación y Evaluación Educativa, 3*(1-2)

Eisner, E. (1998). *El ojo ilustrado: Indagación cualitativa y mejora de la práctica educativa.* Barcelona: Paidós.

Fernández, M., Noelia, A. y Pérez, M. A. (2009). *Curso básico de psicometría: Teoría clásica.* Buenos Aires: Lugar Editorial.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues & Practice, 24*(3), 21-28.

García, S. (2002). La validez y la confiabilidad en la evaluación del aprendizaje desde la perspectiva hermenéutica. *Rev. Ped. 23*(67), 297-318.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report, 8*(4), 597-607.

Gullickson, A. y Joint Committee on Standards for Educational Evaluation. (2003). *The student evaluation standards: How to improve evaluations of students.* Thousand Oaks, CA: Sage.

Hernández, R., Fernández-Collado, C. y Baptista, P. (2007). *Metodología de la investigación* (4a. ed.). México: McGraw-Hill.

Hopkins, K. (1998). *Educational and psychological measurement and evaluation.* Needham Heights, MA: Allyn and Bacon.

Joint Committee on Standards for Educational Evaluation. (2003). *The student evaluation standards.* California: Corwin Press.

Jonson, J. L. y Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58*(5), 736-753.

Lamprianou, I. y Christie, T. (2009). Why school based assessment is not a universal feature of high stakes assessment systems? *Educational Assessment, Evaluation & Accountability, 21*(4), 329-345.

Lewis, R. (1996). *Tests psicológicos y evaluación*. México: Aiken Editorial-Prentice Hall.

Li, H. (2003). The resolution of some paradoxes related to reliability and validity. *Journal of Educational and Behavioral Statistics, 28*(2), 89-95.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. Educational Measurement: *Issues and Practice, 22*(4), 34-43.

Miles, M. y Huberman, M. (1984). Qualitative data analysis: A sourcebook of new methods. EUA: Sage.

Mislevy, R. J. (2004). Can there be reliability without reliability? *Journal of Educational and Behavioral Statistics, 29*(2), pp. 241-244.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), pp. 5-12.

Moss, P. A. (2004). The meaning and consequences of reliability. *Journal of Educational and Behavioral Statistics, 29*(2), 245-249.

Muñoz-Izquierdo, C. (2009). *¿Cómo puede la educación contribuir a la movilidad social. Resultados de cuatro décadas de investigación sobre la calidad y los efectos socioeconómicos de la educación*. México: Universidad Iberoamericana.

Parkes, J. (2007). Reliability as argument. *Educational Measurement: Issues & Practice, 26*(4), 2-10.

Pedhazur, E. y Pedhazur, L. (1991). *Measurement, design and analysis: An integrated approach.* Hillsdale, NJ: LAwrence Erlbaum Associates.

Sadovnik, A. R. (2001). Basil Bernstein (1924–2000). *Prospects, 31*(4), 607-620.

Salkind, N. (2006). *Test & measurement for people who (they think) hate tests & measurement.* Thounsand Oaks, CA: Sage.

Scates, D. E. (1943). Differences between measurement criteria of pure scientists and of classroom teachers. *The Journal of Educational Research, 37*(1), 1-13.

Shepard, L. A., Center for Research on Education, Diversity & Excellence, & University of California, Los Angeles. Center for Research on Evaluation, Standards, and Student Testing. (2000). *The role of classroom assessment in teaching and learning.* Los Ángeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, UCLA.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice, 22*(4), 26-33.

Stiggins, R. J. (1992). High quality classroom assessment: What does it really mean? *Educational Measurement: Issues and Practice, 11*(2), 35-39.

Stiggins, R. J. (1997). Dealing with the practical matter of quality performance assessment. *Measurement in Physical Education and Exercise Science, 1*(1), 5-17.

Teasdale, A. y Leung, C. (2000). Teacher assessment and psychometric theory: A case of paradigm crossing? *Language Testing, 17*(2), 163-184.

Tristán, A. (1998). *Análisis de rasch para todos: Una guía simplificada para evaluadores educativos.* México: Centro Nacional de Evaluación para la Educación Superior.

Winter, G. (2000). A comparative discussion of the notion of validity in qualitative and quantitative research. *The Qualitative Report, 4*(3), 4.