EVALUACIÓN AL INTERIOR DEL AULA: UNA MIRADA DESDE LA VALIDEZ, CONFIABILIDAD Y OBJETIVIDAD

Assessment inside the classroom: a view from the validity, reliability and objectivity approach

Carla Förster Marín*
Cristian A. Rojas-Barahona**

Resumen

El artículo profundiza en el tema de la evaluación en el aula, una práctica cotidiana que pocas veces es analizada. Por medio de elementos conceptuales y experiencias reales de estudiantes se intenta identificar los elementos determinantes y necesarios para una evaluación de calidad. Se analiza la validez de contenido, la validez instruccional, la validez consecuencial, la confiabilidad y la objetividad, desde una mirada psicométrica en pruebas a gran escala y en evaluaciones específicas a nivel de aula. Al mismo tiempo se entregan algunas sugerencias o ideas a tener presentes en la práctica educativa, con el fin de mejorar la evaluación de los aprendizajes, por consecuencia el aprendizaje en sí mismo.

Palabras clave: evaluación de aprendizajes, validez, confiabilidad, objetividad

Abstract

The article deepens into the classroom assessment topic which, despite being a daily practice, is hardly analyzed. By means of concepts and students' real experiences we attempt to identify determinant elements needed for quality assessment. Content, instructional and consequential validity, reliability and objectivity are analyzed from a psychometric view in large scale tests as well as through specific classroom assessment. Additionally, some suggestions worth bearing in mind are provided with respect to improving learning assessment; consequently, learning itself.

Key words: *learning assessment, evaluation, validity, reliability and objectivity*

* Magíster en Evaluación Educacional, Profesora Facultad de Educación, Pontificia Universidad Católica de Chile. Correo: ceforste@uc.cl

** Doctor en Psicología Evolutiva y de la Educación, Profesor Facultad de Educación, Pontificia Universidad Católica de Chile. Correo: crojash@uc.cl

Introducción

Los docentes permanentemente deben emitir un juicio respecto del aprendizaje que alcanzan los estudiantes en las diferentes temáticas que enseñan. Desde la mirada de la evaluación que se presenta en este artículo, este juicio supone que con anterioridad se recogió información de manera sistemática y confiable, se analiza correctamente y se contrasta contra un referente previamente establecido, por tanto, la conclusión obtenida sería válida, confiable y objetiva. Pero, ¿qué ocurre realmente en la sala de clases? Para decir que se va por el camino correcto se debe responder positivamente a las siguientes preguntas: ¿lo que respondieron los estudiantes era lo que quería evaluar el profesor?, ¿las respuestas en la evaluación son coherentes con el dominio de la disciplina que se presenta durante las clases? y ¿los resultados son independientes de quién evalúe y de quién revise la información obtenida?

Las concepciones de evaluación y los enfoques teóricos con que cada docente se aproxima a la evaluación del aprendizaje de sus estudiantes condicionan la forma en que toma decisiones al interior del aula (Brown, 2004; Castillo y Cabrerizo, 2007; Flórez, 1999). Además, hay que tener presente que la información que se recoge siempre está permeada por las características psicológicas, sociales, económicas y físicas, particulares de cada estudiante, y puede estar influida por problemas de índole pasajera, circunstancias especiales del contexto, o situaciones propias de la administración de los instrumentos con los que se recoge dicha información (Luckett y Sutherland, 2000; Salinas, 2002; Sanmartí, 2007). En este sentido, la calidad de una evaluación, es decir, la consistencia, adecuación y pertinencia de los procedimientos e instrumentos que se utilizan para evaluar el desempeño de los estudiantes, debiese considerar la validez, confiabilidad y objetividad como elementos esenciales al momento de interpretar los resultados y así tomar decisiones adecuadas respecto de qué y cómo enseñar (Brookhart, 2003; Himmel, Olivares y Zabalza, 1999; McMillan, 2003).

El presente artículo tiene como objetivo analizar desde una perspectiva teórico-práctica los conceptos de validez, confiabilidad y objetividad de una evaluación, y proponer sugerencias que permitan a los docentes resguardar estos criterios en futuras evaluaciones de aprendizajes al interior de sus aulas.

Validez

El primer elemento a considerar en la calidad de una evaluación es la validez de la información recogida con un instrumento o en una situación evaluativa. Este concepto tiene su origen en la medición y se ha desarrollado ampliamente desde el enfoque psicométrico, donde se dice que una prueba es válida "si mide lo que pretendía medir"; no obstante, esta definición genérica debe complementarse con la finalidad para la que fue

construida, es decir, si la interpretación de una determinada puntuación permite señalar correctamente una conclusión respecto del fin o constructo que mide dicha prueba (Hogan, 2004). Concordante con esta idea, Messick (1989) define la validez como "*un juicio evaluativo integrado del grado en el cual la evidencia empírica y teórica soporta la adecuación y pertinencia de las inferencias y acciones basadas en los puntajes de un test o en otros modos de evaluación*" (p. 13). Así, por ejemplo, la pregunta ¿la Prueba de Selección Universitaria (PSU) es válida? no es una buena pregunta, más bien correspondería decir: ¿la PSU es válida para predecir los resultados de los estudiantes durante su primer año de universidad?, o mejor aún, ¿en qué grado es válido el puntaje de la PSU para predecir el resultado de los estudiantes en su primer año de universidad? En otras palabras, la validez no es una propiedad intrínseca de un instrumento o situación evaluativa, sino una propiedad de la interpretación y los usos que se realizan de la información obtenida a partir de ellos (Valverde, 2000).

El concepto de validez utilizado en el enfoque psicométrico tiene sentido solo para pruebas a gran escala, aplicadas a un número considerable de estudiantes, donde se espera que los resultados generen una distribución normal (se espera que la mayoría de las personas obtenga puntuaciones medias y una proporción menor se ubique en los extremos), pero en la sala de clases no se tiene como objetivo que los resultados de los alumnos se distribuyan de esta forma, por el contrario, se espera una curva asimétrica, donde la mayoría (si no todos) esté sobre el puntaje de aprobación de un determinado aprendizaje, ya que esto indicaría que lograron aprender lo que se esperaba para ese momento (edad, curso, período del año, etc.). Por esta razón, Stiggins (2001) evita hablar de "validez" en el contexto de la sala de clases y reemplaza este término por evaluación de "alta calidad" (*high-quality classroom assessment*).

Diversos autores plantean que en la sala de clases el significado de validez no es el tradicional, ya que la evidencia estadística es muy difícil de obtener (Luckett y Sutherland, 2000; McMillan, 2003; Moss, 2003; Stiggins, 2001). Más aún cuando la mayoría de los profesores no tiene conocimientos en teoría de la medición, producto de la escasa presencia de esta disciplina en su formación inicial y, por tanto, muy pocas herramientas para hacer adaptaciones y reconocer las limitaciones de la "validez psicométrica" en la sala de clases (Brookhart, 2003). Asimismo, Moss (2003) se pregunta hasta qué punto la utilización de la validez propia de la medición psicométrica asegura una orientación hacia las temáticas relevantes en la sala de clases y cómo aportarían otras perspectivas teóricas a generar un marco sólido para obtener evaluaciones válidas al interior del aula.

Brookhart (2003), por medio de un cuadro comparativo, realiza una clara diferenciación de la validez de una evaluación según el contexto para el cual fue diseñada (ver Tabla 1). Destaca principalmente el propósito de la validez y la importancia del contexto de aplicación de la evaluación. A gran escala el propósito es sacar conclusiones del desempeño de los estudiantes para los fines específicos de la evaluación diseñada,

mientras que a nivel de aula es determinar el desempeño de los estudiantes y entregar evidencia respecto de la enseñanza de ese contenido. Con respecto al contexto, las evaluaciones a gran escala no la consideran como una variable relevante, mientras que a nivel aula sí es un tema a considerar, debido que importan, por ejemplo, los conocimientos previos del alumno y su contexto sociocultural.

Tabla 1
COMPARACIÓN DE LA VALIDEZ EN EVALUACIONES
A GRAN ESCALA Y EN LA SALA DE CLASES
(Adaptado de Brookhart, 2003)

Evaluación a gran escala	Evaluación a nivel aula
Las interpretaciones y las acciones tomadas son externas al proceso de medición.	Las deducciones y las acciones tomadas son internas al proceso de medición.
Los estudiantes son la "materia" sobre la que se realizan las observaciones, no tienen injerencia en la evaluación.	Los estudiantes son observadores conjuntamente con los profesores; "esas mediciones" corresponden a procesos formativos de evaluación.
Los estudiantes no reciben un beneficio directo de la evaluación, los análisis no son individuales.	Los estudiantes reciben beneficios individuales de la información de la evaluación.
El objetivo de la evaluación es realizar una inferencia significativa estadísticamente sobre el rendimiento de los estudiantes y/o la utilización eficaz de esa información para una finalidad determinada.	El objetivo de la evaluación es comprender cómo es el desempeño de los estudiantes comparado con el "ideal" (definido en los objetivos de aprendizaje) y/o el uso efectivo de esa información para el aprendizaje.
El contexto de la medición es un constructo irrelevante.	El contexto de la medición es un constructo relevante.
Las especificaciones del contenido describen un dominio disciplinar.	Las especificaciones de contenido reflejan el dominio (objetivo de aprendizaje) y la enseñanza (modos, actividades).
La administración del test es estandarizada y se controla el efecto de las ciencias, prácticas y diferencias culturales.	Las creencias de los profesores, sus prácticas y el conocimiento de la materia y de los estudiantes (incluyendo diferencias culturales y lingüísticas) son preocupaciones relevantes.
La evaluación es independiente de la enseñanza. Se puede relacionar con diferentes contextos a través de cruces de datos.	La evaluación es parte de la enseñanza. Una buena evaluación es un "episodio genuino de aprendizaje".

Aunque es claro el hecho de que la validez de una evaluación en la sala de clases no obedece a una lógica psicométrica (Brookhart, 2003), es posible rescatar algunos elementos de los distintos tipos, categorías o aspectos relacionados con la validez que se han descrito en diversas clasificaciones, donde las más clásicas son la validez de contenido, de constructo y la relacionada a criterio (Gorin, 2007; Hogan, 2004; Moss, 2007; Sireci, 1998). Lo primero que se observa al introducirse en los textos especializados es que hasta el día de hoy no existe un acuerdo en la comunidad científica respecto de si hay una validez (la de constructo) que incluye a las otras o si son de naturaleza diferente y excluyente, y por tanto estar separadas (Borsboom, Mellenbergh y van Heerden, 2004; Gorin, 2007; Moss, 2007; Sireci, 1998). En este sentido, se rescatan las palabras de Valverde (2000): "*Existe una gran cantidad de opciones en cuanto al tipo de evidencia que se puede acumular y reportar. Cada tipo de evidencia ilumina y da apoyo a distintas facetas de la validez*" (p. 26). Así, en el presente artículo no se pretende profundizar en la discusión ni adherir a una clasificación en particular, sino considerar cuáles de estas categorías o aspectos serían relevantes de considerar para resguardar la calidad de la evaluación en el contexto educacional, específicamente en la sala de clases. A continuación se presenta una revisión conceptual de las tipologías relevantes en la evaluación al interior de la sala de clases y se indican algunos ejemplos de situaciones de evaluación descritas por los estudiantes, en las cuales se evidencia la falta de validez.

Validez de contenido

La validez de contenido es un concepto que ha resultado controversial a través del tiempo (ver resumen de la evolución del concepto en Sireci, 1998). Teóricos como Messick (1989) señalan que es parte de la validez de constructo y que dada su naturaleza "cualitativa" es técnicamente incorrecto y no corresponde a una nomenclatura psicométrica, proponiendo hablar de relevancia, representación o cobertura de los contenidos. Gorin (2007), por su parte, señala que un constructo se mantiene independiente del contexto de evaluación, mientras que un contenido es dependiente de dicho contexto. Hogan (2004) plantea que para determinar la validez de constructo de un instrumento se requiere analizar su *estructura interna*[1] y *los procesos de respuesta*. Como ambos procedimientos son muy difíciles de llevar a cabo en el aula y el contexto es un elemento importante que condiciona las situaciones de evaluación, se presenta la validez de contenido como un elemento clave a considerar para resguardar la calidad de una evaluación.

[1] Consistencia interna o confiabilidad de dicho instrumento, la cual se calcula estadísticamente con KR-20 o alfa de Crombach y las dimensiones comunes que subyacen al modelo a través de análisis factorial.

Como ya se mencionó, no es tema de este artículo discutir si es una categoría diferente o parte de la validez de constructo, sino rescatar la importancia de considerar este criterio para resguardar la calidad de una evaluación. Esta "validez" se refiere a la correspondencia que existe entre el contenido/habilidades que evalúa el instrumento y el campo de conocimiento al cual se atribuye dicho contenido (Brualdi, 1999; García, 2002; Hogan, 2004; Lukas y Santiago, 2004). En el ámbito educacional se ha trabajado ampliamente en generar índices de validez de contenido en pruebas a gran escala, considerando variables como: a) la opinión de expertos (profesores, investigadores, curriculistas, directivos) en cuanto a la cobertura curricular de las pruebas y la cobertura instruccional de dichos contenidos, b) la muestra de estudiantes a quienes se les aplicó la prueba, c) el número de ítems de cada contenido, entre otras (Crocker, Llabre y Miller, 1988). Sin embargo, estos modelos estadísticos diseñados para calcular índices de validez de contenido no son aplicables a las situaciones evaluativas que se realizan al interior del aula, dado que no cumplen con los requisitos de base que sostienen dichos modelos.

En este sentido, resguardar que las evaluaciones en el aula tengan validez de contenido implica garantizar que las situaciones evaluativas incluyan los contenidos y habilidades asociadas a estos. Se deben considerar los contenidos y habilidades más relevantes para demostrar el logro de un aprendizaje, es decir, se debe seleccionar una muestra pertinente y representativa del universo de contenidos cubierto por un curso, una clase, o un conjunto de ellas. Sireci (1998), luego de una exhaustiva revisión bibliográfica, plantea que hay consenso en al menos cuatro elementos que caracterizan el concepto de validez de contenido: la definición de un dominio disciplinar, la relevancia de dicho dominio, la representatividad del dominio y los procedimientos de construcción del instrumento o situación evaluativa. Esto es particularmente importante porque las conclusiones que se pueden sacar a partir de una evaluación solo son válidas para lo que se evaluó, es decir, los puntajes o categorías que obtenga cada estudiante deben provenir de tareas y estímulos de calidad, coherentes con los propósitos conceptuales para los que fueron elaborados. Por tanto, certificar lo que un estudiante sabe y es capaz de hacer o no en un curso tiene directa relación con los contenidos y habilidades que fueron evaluados.

A continuación se exponen algunos ejemplos de evaluaciones consideradas negativas por los estudiantes y que tienen relación con la falta de validez de contenido:

- *"Durante un curso se dio la situación de que el profesor dio bastante lectura para una prueba y en ella comenzó a preguntar datos demasiado específicos, palabras exactas citadas en los textos para identificar de qué autor podían ser y nombres y fechas, que si bien eran interesantes o importantes, para una evaluación de tantos contenidos eran muy específicos".*
- *"Para mí son negativas todas aquellas evaluaciones en las que se preguntan citas textuales de la bibliografía, que no están relacionadas con el marco teórico general del curso".*

- *"Recuerdo una prueba que involucraba muchos contenidos pero que nos habían anticipado no sería memorística, sin embargo, a la hora de enfrentarme a la prueba era evidente que las respuestas debían ser literales, pues las preguntas eran muy específicas y no apuntaban a lo que en el curso se había considerado como relevante".*
- *"Me ha tocado responder evaluaciones donde las preguntas no tienen nada que ver con los contenidos que entraban, parecieran evaluar cualquier cosa menos lo que debiesen".*
- *"Creo que una situación evaluativa negativa podría ser el caso de matemáticas, cuando por ejemplo nos ponían unos ejercicios que se podían resolver solo si se te ocurría el truco adecuado. Creo que es negativo porque el aprender no depende de si te sabes o no el truco, sino de resolver y plantear bien un problema".*
- *"Hay veces en que tenemos que estudiar, por ejemplo un texto entero, y luego en la prueba aparecen un par de preguntas textuales y nada más… me parece negativo porque creo que es bueno aplicar lo que uno lee, además suelen pasar que justo nos preguntan lo que teníamos menos claro de todo y eso nos perjudica".*
- *"En una clase nos pidieron hacer unos gráficos de la economía de la región y luego el profesor evaluó más la estética, que si estaban o no bien elaborados, y los compañeros que pintaron los gráficos de colores les fue mejor que a mi grupo que teníamos los gráficos en blanco y negro… ahí aprendí a preocuparme más por la estética que por los contenidos".*

Como se puede apreciar en estos relatos, los estudiantes resienten que se les evalúen contenidos demasiado específicos, carentes de significado o que no tienen relación con el dominio disciplinar esperado.

A continuación se presentan algunas sugerencias para resguardar la validez de contenido de una situación evaluativa, que corresponden a procedimientos y acciones que un profesor de aula podría llevar a cabo en la elaboración de sus propias evaluaciones. Si bien hay sugerencias en el ámbito estadístico que también se pueden realizar, no han sido consideradas en este análisis:

1) *Elaborar una tabla de especificaciones del instrumento*, que considere: el contenido, la meta/objetivo/indicador de aprendizaje/logro específica/o que se desea abordar con ese contenido, la habilidad cognitiva, afectiva o psicomotora que se desea evaluar asociada a ese contenido y los ítems que se han elaborado para ello. Con esta tabla es posible visualizar si se están cubriendo todos los contenidos seleccionados, si las habilidades asociadas a dichos ítems son coherentes con las metas de aprendizaje trazadas y si los ítems corresponden a los contenidos y habilidades que se esperaban. A partir de este análisis, también es posible establecer si existe una graduación de la complejidad en cuanto a las habilidades evaluadas y si los ítems son los más apropiados para evaluar una habilidad determinada.

2) *Verificar a través de criterio de jueces que la situación de evaluación sea adecuada para medir los contenidos planteados en el aprendizaje*. Dado que en la sala de clases es el mismo profesor quien construye los instrumentos y situaciones evaluativas, muchas veces se pasan por alto aspectos que para quien no ha visto antes dicha evaluación son más evidentes. Por esta razón se recomienda mostrar a otros profesores el instrumento con su pauta de corrección y la tabla de especificaciones para que den su opinión. Puede ser que el problema se encuentre en la redacción de una instrucción o en la pauta de respuesta, y al mirar solo la tabla esto no es evidente.

Validez instruccional

Según Hogan (2004), esta validez corresponde a una aplicación particular de la validez de contenido y es conocida también como validez curricular. Tiene relación con lo que los estudiantes han tenido oportunidad de aprender durante las clases para responder correctamente en una evaluación (Crocker *et al.*, 1988).

En el ámbito educativo este tipo de validez es clave, dado que representa la relación entre lo que se enseña y lo que se evalúa. Cuando esta relación es débil se presentan dos problemas, por una parte, los estudiantes no tienen posibilidad de demostrar lo que aprendieron durante las clases y, por otra, son evaluados en aspectos que no se les enseñaron (Himmel *et al.*, 1999; McMillan, 2003). Esta última idea se ve reflejada especialmente cuando se cambia el énfasis de lo que se enseñó, por ejemplo, en clases se enseñan los conceptos, sus definiciones y luego en la evaluación se les pide que apliquen dichos conceptos en situaciones que nunca han sido trabajadas durante las clases, aludiendo a que se espera que los estudiantes sean capaces de hacerlo como parte de la "construcción" de su propio aprendizaje.

En evaluaciones a gran escala se incorporan otros problemas éticos relacionados con la interpretación que se hace de los resultados: responsabilizar a los estudiantes por no saber contenidos que no han tenido la oportunidad de aprender y responsabilizar a los docentes por el bajo logro de sus estudiantes sin proporcionarles las condiciones, materiales o la capacitación necesaria para enseñar los contenidos (Valverde, 2000). En conclusión, se dice que una evaluación tiene validez instruccional cuando contiene situaciones evaluativas coherentes con las actividades de aprendizaje realizadas por los alumnos.

En este apartado se incluye también lo que se ha llamado "validez semántica", que consiste en que las situaciones de evaluación contienen términos cuyo significado es conocido y compartido entre el constructor del instrumento (en el caso del aula, el profesor) y los alumnos. Muchas veces se enseña un contenido usando un término y

luego en la evaluación se pregunta utilizando otro término "sinónimo" que no ha sido trabajado. En estricto rigor, el contenido fue enseñado, pero frente a una respuesta errónea de los estudiantes cabe preguntarse si no dominan el contenido o fue la palabra que no comprendieron lo que provocó dicha respuesta.

A continuación se presenta una serie de situaciones relatadas por los estudiantes que resultaron negativas para ellos y que aluden a la validez instruccional.

- *"Ha sido negativo para mi aprendizaje el hecho de que muchas veces los profesores evalúan materia o contenidos que no correspondía a lo que habíamos visto en las clases. Eso me acarreó a la larga una desmotivación y frustración al no conseguir la calificación esperada o al no poder aplicar lo que había estudiado".*
- *"Fue negativo para mí cuando las clases teóricas no iban acordes a las evaluaciones de laboratorio y nos evaluaban materia que aún no habíamos visto, hay muchas actividades de laboratorio que aún no logro comprender del todo".*
- *"Las pruebas en un curso no correspondían a lo que el profesor enseñaba en clases, enseñaba mal, creaba un clima de terror en la sala y luego nos preguntaba cosas que no había enseñado o que había pasado en el curso del lado".*
- *"En un ramo de mi carrera, la primera prueba que nos hicieron no correspondía con lo que trabajábamos y aprendíamos en clases. Solo hacíamos teoría y la prueba fue de aplicación, con muchos detalles, y el profesor no hacía ni destacaba esos elementos. En suma, lo que aprendí no me sirvió para la prueba".*
- *"He tenido profesores que evalúan en sus pruebas cosas (contenidos) que han tratado mal, a la rápida o sencillamente no han tratado ni en clases ni en los textos referidos (asumiendo que los vimos en otros cursos). No me parece pedagógico evaluar contenidos de otros cursos".*

Como se puede observar las opiniones de los estudiantes aluden directamente al desempeño profesional del docente, por tanto, las acciones que se sugieren a continuación para resguardar la validez instruccional tienen relación con las prácticas de enseñanza que se utilizan habitualmente:

1) *Velar porque las situaciones de evaluación contengan los contenidos vistos en las actividades de aprendizaje realizadas*. Esto supone llevar un registro de lo que se enseñó en cada clase, para elaborar las situaciones de evaluación no basta con utilizar la calendarización o planificación, a menos que ésta sea actualizada permanentemente. Muchas veces las dinámicas propias de un curso permiten profundizar o volver atrás en una temática, o sucede alguna actividad no planificada, produciéndose desfases temporales que pueden hacer que se evalúen contenidos que aún no han sido revisados o que se vieron en un curso, pero en otro aún no.

2) *Velar porque las situaciones de evaluación sean equivalentes o similares a las actividades de aprendizaje realizadas*. No significa repetir la guía que se realizó en clases y ahora ponerle una nota, por el contrario, esta sugerencia apunta a evaluar actividades en término de habilidades similares a las trabajadas. Esto supone que durante el proceso de enseñanza se intencionaron actividades para lograr los aprendizajes esperados o metas de aprendizaje determinadas para esos estudiantes y, por tanto, se debe evaluar si se lograron o no. Se sugiere, por una parte, hacer uso de la planificación de clases y que esta sea una guía flexible de apoyo a la preparación de la enseñanza; y, por otra, utilizar la tabla de especificaciones, poniendo énfasis en que las situaciones de evaluación sean coherentes con lo que cada profesor trabajó con sus estudiantes. En este último punto, los "jueces expertos" no tienen competencia ya que no estuvieron durante el proceso, por lo que solo depende de la autoevaluación que realice el profesor respecto de su propia práctica.

3) *Cuidar que el lenguaje utilizado en las situaciones de evaluación sea conocido por los estudiantes*. Esto implica tener especial cuidado en utilizar términos familiares para los estudiantes, tanto en las instrucciones como en el lenguaje técnico propio de la disciplina que se desea evaluar. Es importante hacer notar que el fin de una evaluación es obtener información del aprendizaje de los estudiantes y las "palabras truco" sólo permiten generar información de mala calidad respecto de ese fin.

Validez consecuencial

Este tipo de validez se relaciona con las consecuencias y secuelas intencionales y no intencionales que tendrá el uso e interpretaciones que se dará a la información recogida en la evaluación (Hogan, 2004; McMillan, 2003; Moss, 1997). Aunque no hay acuerdo entre los autores respecto de si esta validez es pertinente al ámbito psicométrico (Borsboom *et al.*, 2004; Hogan, 2004; McMillan, 2003), a opinión de Moss (2003) y McMillan (2003) debería ser la principal consideración al momento de tomar decisiones respecto de una evaluación en la sala de clases.

Los aspectos consecuenciales de la validez son especialmente importantes en una evaluación cuando las interpretaciones de la información pueden implicar consecuencias adversas para los participantes (Brualdi, 1999), por ende, estudiar la validez de las consecuencias podría ayudar a controlar estos aspectos. En las evaluaciones a gran escala las consecuencias están en la toma de decisiones y reformulación de políticas públicas (Schutz & Moss, 2004) y tienen implicancias éticas y sociales (Borsboom *et al.*, 2004) que hacen relevante estudiar la validez desde este punto de vista. En la sala de clases, en cambio, las consecuencias tienen relación directa con las dinámicas de enseñanza-aprendizaje que se dan entre profesores y alumnos; en este sentido, Brookhart (2003)

señala que si la integración entre enseñanza y evaluación es tomada seriamente, una situación evaluativa debería ser como consecuencia un vehículo para el aprendizaje.

Por tanto, se podría decir que la validez consecuencial a nivel de aula tiene relación directa con los efectos de la evaluación sobre la enseñanza y los aprendizajes de los estudiantes (Himmel *et al.*, 1999), por consecuencia, se relaciona con los propósitos para los cuales se diseñó la evaluación. Así, McMillan (2003) plantea una serie de preguntas que, de ser respondidas positivamente, indicarían que una evaluación tiene validez consecuencial: ¿Los estudiantes adquieren una comprensión profunda de lo que están aprendiendo mientras se preparan para la evaluación? ¿Los estudiantes creen que son capaces de aprender nuevos conocimientos después de autoevaluarse en ejercicios de práctica de una evaluación? ¿Los estudiantes son capaces de transferir sus conocimientos en las tareas siguientes? La decisión de llevar a cabo revisiones adicionales, como evaluaciones formativas, ¿llevan a un mayor aprendizaje del estudiante?

Estas preguntas apuntan a propósitos diferentes y no siempre se darán todas al mismo tiempo, lo importante es tener en cuenta que la evaluación en la sala de clases siempre tiene consecuencias y, por tanto, los docentes deben pensar en ellas antes, durante y después de realizar la evaluación, para sacar el mayor provecho de esta instancia y minimizar los efectos negativos que se puedan generar. Entre las consecuencias imprevistas de una evaluación, y que le restan validez, destacan la falta de motivación de los estudiantes para participar en una actividad de evaluación, interiorizar un error conceptual a causa de la tarea realizada, generar un ambiente de competencia entre los estudiantes, resentimiento de los estudiantes por el formato de la evaluación, entre otros. (Brookhart, 2003; McMillan, 2003).

Los estudiantes relatan una serie de situaciones que para ellos fueron negativas, donde se evidencia la validez consecuencial:

> *"En un trabajo en grupo, en el cual se evaluó a todos por igual, uno de mis compañeros no se preparó bien, se puso nervioso y la exposición no fue la mejor, pero los otros tres compañeros hicieron su presentación de buena forma. Por las razones anteriores, la evaluación del grupo fue disminuida y con esto se premió al compañero deficiente y se perjudicó a los otros que sí lo hicimos bien".*

> *"Un ejemplo de situación negativa correspondería a evaluaciones que se basan en la amenaza de sanciones en caso de no ser aprobadas. Si bien podría otorgar una presión que provoque mayor conciencia y en consecuencia mayor efectividad, a juicio particular no es correcto provocar miedo o algún tipo de daño psicológico al grupo evaluado".*

> *"Una evaluación oral, para mí es muy confuso dar a conocer lo que se sabe con tanta presión de por medio, los nervios juegan en contra en estos casos y siento que no representa una buena manera de evaluar ya que lo que preguntan se podría responder perfectamente en forma escrita y lo hacen más por no revisar".*

"Recuerdo pruebas demasiado extensas en que las preguntas eran exclusivamente de memorización de contenidos y no había oportunidad de pensar y analizar nada. Me aprendía las cosas y me fue bien, pero no me acuerdo de nada, es como si no hubiese hecho el curso y ahora lo voy a necesitar para mi futuro laboral".

Puesto que las consecuencias de una evaluación pueden ser negativas, y de largo plazo en nuestros estudiantes, las sugerencias que se dan para resguardar la validez consecuencial se basan en su mayoría en acciones relacionadas con la definición y planificación de una evaluación:

1) *Definir claramente los propósitos y usos de la evaluación*. Como esta validez se basa en las implicancias que tiene un proceso evaluativo para los estudiantes, es importante definir previamente la finalidad que tiene realizarlo, es decir, ¿para qué evaluamos? En este sentido se presentan tres grandes fines: diagnóstico (por ejemplo, identificar los conocimientos previos de los estudiantes, reconocer conductas de entrada, ajustar la planificación al contexto, etc.), formativo (por ejemplo, mejorar los aprendizajes y la motivación de los estudiantes, monitorear su desarrollo, mejorar el proceso de enseñanza, etc.) y sumativo (por ejemplo, certificar los aprendizajes de los estudiantes, informar a los padres el desempeño de sus hijos al final del curso, etc.), a partir de los cuales se tomarán decisiones que tendrán implicancias para la enseñanza y el aprendizaje de los alumnos. Es importante tener presente que un mismo instrumento puede ser válido consecuencialmente para una instancia y no tener ninguna validez en otra.

2) *Identificar claramente las evidencias que darán cuenta de los propósitos de la evaluación*. Como ya se ha señalado anteriormente, la validez consecuencial está determinada por la correspondencia entre la información que se recoge y los fines para los que se utilizará. En este sentido, la sugerencia es tener una buena planificación que permita resguardar esta correspondencia y no realizar una evaluación y luego definir para qué la puedo usar. Un ejemplo de esto son las "pruebas sin aviso" que se realizan en una clase, cuya motivación obedece más a un control disciplinario que a la evaluación de los aprendizajes esperados. Estas evaluaciones por lo general son improvisadas y se agregan como una calificación cuyas consecuencias son, en su mayoría, negativas para los estudiantes.

Confiabilidad

Otro elemento que se considera al momento de analizar la calidad de una evaluación es la confiabilidad o precisión de ésta. A diferencia de la validez, la confiabilidad solo se relaciona con la consistencia de la medición, al margen de lo que se mida exactamente (Flórez, 1999; Hogan, 2004). La confiabilidad implica que el instrumento entrega resultados similares cuando se repite su aplicación en las mismas circunstancias a las

mismas personas. El concepto de confiabilidad hace referencia a consistencia, exactitud y estabilidad de los resultados, a las inferencias que se pueden realizar, y tiene directa relevancia en las conclusiones y posterior toma de decisiones (Luckett & Sutherland, 2000; McMillan, 2003). La confiabilidad se puede verificar de diversas maneras, pero las más comunes son el coeficiente alfa de Crombach, que entrega la consistencia interna de un instrumento, la correlación test-retest, que evalúa la consistencia entre dos mediciones del mismo test, aplicadas al mismo sujeto, y las formas paralelas, que evalúan el grado de correlación que hay entre dos versiones de una misma prueba.

Este criterio de calidad tiene su origen en la psicometría, donde se dice que una prueba será confiable si consistentemente genera la misma puntuación, o una similar, al ser aplicada a un individuo, y su puntuación es replicable al menos con un margen de error pequeño. Moss (2003) señala que en las evaluaciones de aula, a diferencia de las evaluaciones a gran escala, la confiabilidad no es un tema relevante y que muy pocas veces se considera; sin embargo, argumenta que su afirmación se refiere a pruebas, y que en la sala de clases los profesores tienen múltiples instancias para recoger información complementaria de un aprendizaje. Smith (2003) reafirma esta idea señalando que los profesores no calculan coeficientes alfa, ni correlaciones test-retest, ni formas paralelas, por una parte, porque no tienen un número de casos suficientes y, por otra, porque las evaluaciones son realizadas a un estudiante en un tiempo específico. No obstante, lo que es fundamental en este caso es que se espera que el alumno cambie su aprendizaje de una semana a otra, elemento que va contra lo esperado en psicometría respecto de la estabilidad de los rasgos que se miden.

Por tanto, en el contexto de la sala de clases, donde no interesa generar un orden o jerarquía de los estudiantes en las actividades que se evalúan, la confiabilidad desde una mirada psicométrica no es pertinente (Brookhart, 2003; Smith, 2003). Además, muchas evaluaciones se califican con categorías (por ejemplo, excelente, satisfactorio, regular, insatisfactorio), lo que generaría un problema para hacer algún cálculo de confiabilidad "tradicional". Moss (2003) plantea que lo correcto para hablar de confiabilidad a nivel de aula es analizar la "suficiencia de información", es decir, la confiabilidad está representada por tener suficiente evidencia de un aprendizaje que permita tomar decisiones con el menor margen de error. En su propuesta, se plantea que esta evidencia debe no solo corresponder al mismo contenido sino tener un mismo nivel de exigencia, lo que parece ser lo más difícil de cumplir (Smith, 2003). Brookhart (2003) es muy claro en el momento de diferenciar la confiabilidad a gran escala y a nivel de aula (ver Tabla 2). La confiabilidad a gran escala se entiende como la consistencia entre los factores que componen un instrumento de medición, mientras que a nivel de aula es vista como suficiencia de la información. En el primer caso se busca una clasificación estable a lo largo de un continuo y a nivel de aula se busca obtener información estable entre el ideal del aprendizaje (lo que busca el profesor de sus alumnos) y lo real alcanzado.

Tabla 2
COMPARACIÓN DE LA CONFIABILIDAD DE EVALUACIONES
A GRAN ESCALA Y EN LA SALA DE CLASES
(Adaptado de Brookhart, 2003)

Evaluación a gran escala	Evaluación en el aula
La confiabilidad es la consistencia entre factores relevantes.	La confiabilidad es la suficiencia de la información.
El objetivo de la confiabilidad es tener una clasificación estable de los estudiantes en una escala de puntuación o una categorización estable a lo largo de un continuo de progreso.	El objetivo de la confiabilidad es tener información estable acerca de la diferencia entre el desempeño de los estudiantes y el "ideal", tal como se definen en los objetivos de aprendizaje.

Entre los factores que pueden influir en la confiabilidad de una evaluación destacan: a) el número de observaciones o evidencias de un aprendizaje, ya que mientras más instancias de evaluación del mismo aprendizaje (número de ítems en una prueba, instancias de evaluación formativas, etc.), mayor será la "consistencia interna" del proceso de evaluación y más confiable será la conclusión respecto del logro del estudiante; b) las características de la aplicación, referidas a la claridad de las instrucciones, tiempo destinado y disposiciones del espacio físico; y c) la precisión de la corrección y puntuación (Himmel *et al.*, 1999; Hogan, 2004). Este último punto será tratado en forma independiente bajo el título de "Objetividad", dada la relevancia que tiene en el ámbito educativo.

Los estudiantes frente a este criterio de calidad señalan que sus malas experiencias evaluativas se centran principalmente en la falta de oportunidades para demostrar su aprendizaje en distintas instancias. Algunos ejemplos se presentan a continuación:

- *"En un taller en la universidad, solo había una prueba al final y con eso se jugaba la nota del curso. Encuentro que eso es una mala evaluación, porque hay muchas cosas que pueden afectar la respuesta que uno da y justo ese día puede pasar algo que haga que uno responda mal, aunque haya estudiado".*
- *"Una situación negativa de evaluación fue cuando tenía que disertar y estaba enferma con fiebre y me sentía muy mal; la profesora me retó porque mi exposición fue muy mala. Me dio mucha rabia, porque yo me sabía la materia y no tuve otra oportunidad de demostrarlo".*
- *"En un curso donde las pruebas eran una pregunta por cada tema y si justo no te sabías eso específico, perdías con esa pregunta, aunque te supieras el tema".*

- *"En las pruebas de alternativas de una profesora, donde en general la alternativa más larga era la correcta, y entonces, por descarte uno podía responder las pruebas y sacar buena nota, al final creo que no aprendí casi nada".*

Para resguardar que las evaluaciones sean confiables se debe tener en cuenta que el fin último de este criterio es que las conclusiones y posteriores decisiones respecto de la forma que se aborda el proceso de enseñanza-aprendizaje con los alumnos debe estar basado en la evidencia suficiente para no cometer errores (o al menos minimizarlos). En este sentido las acciones sugeridas a continuación proporcionan una guía para lograr evaluaciones confiables:

1) *Aplicar al mismo estudiante variadas situaciones evaluativas que midan el mismo aprendizaje*. Dar la posibilidad al estudiante de demostrar su desempeño frente a un mismo aprendizaje, a través de reiteradas oportunidades, permite tener la certeza de que el aprendizaje fue o no logrado. Esto se puede realizar en distintas actividades evaluativas o poniendo en un mismo instrumento distintas situaciones de evaluación que evalúen lo mismo (por ejemplo, distintos ítems en una prueba).

2) *Velar por la claridad de los ítems e instrucciones*. Hay ocasiones en que un ítem puede estar correctamente construido, pero ser comprendido por los estudiantes en dos sentidos diferentes, si no está esto considerado, al momento de revisar la información se puede generar un sesgo que altere las puntuaciones de los estudiantes y, por tanto, la confiabilidad de la evaluación.

3) *Velar porque el ambiente de aplicación sea similar en cuanto a recursos, espacios y tiempo*. Es particularmente importante tener en cuenta esta sugerencia cuando se realizan evaluaciones en ambientes que no son los óptimos, por ejemplo: hay música en el patio, hace mucho calor en la sala, hay una actividad recreativa programada a continuación, etc., pues los resultados de la evaluación estarán influidos por estos aspectos y puede que no representen los aprendizajes reales de los estudiantes.

4) *Velar por la precisión de la corrección*. Esta sugerencia se refiere a la objetividad de la revisión de la información recogida en la evaluación y será tratada con detalle a continuación.

Objetividad

La objetividad o precisión de la corrección en un proceso evaluativo es un elemento clave asociado a la confiabilidad de una evaluación; sin embargo, se decidió tratarlo en forma separada dada la relevancia e impacto que tiene al interior de la sala de clases. Esto no supone que sea independiente de la confiabilidad.

Por lo general, la objetividad se entiende como la calidad de un objeto en sí, independiente de las consideraciones o juicios personales. Si llevamos la objetividad

al ámbito evaluativo, supone que tanto los instrumentos como el juicio que se emite a partir de la información recogida con ellos sean imparciales. En relación a este tema Calatayud (1999) afirma que quien cree que la evaluación de los estudiantes es una acción objetiva, se embarca en una tarea imposible. La autora plantea que la evaluación es una práctica compleja, que involucra no solo el dominio de una técnica, sino también posee una carga moral y valórica importante, y por tanto, es una actividad ligada a las creencias personales de los docentes, lo quieran o no. Desde esta perspectiva, la imparcialidad no es posible y cada juicio que se emite tiene un componente de apreciación propio de quien lo expresa.

Esta preocupación por la objetividad de la evaluación no es un tema nuevo. Estudios asociados a las expectativas de los docentes como los de Rosenthal y Jacobson (1968; efecto "Pigmalion"), de Spears (1984) y de Rubie-Davies (2006) muestran que los estudiantes considerados de "bajo rendimiento" tienen menos tiempo para contestar en una interrogación oral, reciben menos ayuda o "pistas" cuando preguntan alguna duda en una prueba y reciben menos retroalimentación de su desempeño que los estudiantes considerados de "alto rendimiento", para los cuales los docentes tienen mayores expectativas. Cabe señalar que estas expectativas, en general, están condicionadas por características personales de los estudiantes como el género, el nivel socioeconómico, la raza o etnia, la apariencia física y los patrones de lenguaje oral.

Estudios como el de López y colaboradores (1983) dan cuenta de cómo diferentes profesores de matemáticas o física dan puntuaciones o calificaciones distintas a un mismo ejercicio, lo cual se podría atribuir a diferentes niveles de exigencia en la corrección que no debería afectar a los estudiantes, si son corregidos todos por la misma persona (Gil Pérez y Vilches, 2008), mientras que otras evidencias indican que un mismo profesor puede puntuar o calificar de manera diferente la respuesta de un estudiante dependiendo del momento (Hoyat, 1962). Aunque se asume que la corrección nunca será ciento por ciento objetiva, es conocido por todos que los profesores tienen escaso tiempo para corregir trabajos y pruebas de sus estudiantes, y que lo hacen en los pocos espacios físicos y temporales que tienen disponibles, implicando que algunos comienzan en el colegio durante la mañana, pero pueden terminar al día siguiente en su casa de madrugada. Si no se generan acciones concretas que eviten distorsionar el criterio de corrección, las posibilidades de cambiar la exigencia y la respuesta esperada pueden ser enormes.

Estos problemas de objetividad se ven reflejados en las opiniones de estudiantes respecto de experiencias de evaluación que resultaron ser negativas para su proceso de aprendizaje:

- *"En un taller se hacían controles de lectura con tres preguntas sin mayor explicación, al tener el control observé que no tenía todo el puntaje en ninguna de las preguntas, al preguntar por qué, el profesor me dio la respuesta correcta y era lo mismo pero con otras palabras"*.

- *"Una prueba en un ramo donde casi todo el curso se sacó un rojo. Nadie entendía por qué estaba todo mal. Las preguntas fueron muy específicas, y la pauta era muy estricta. No quedó claro qué estaba malo y qué era correcto".*
- *"En la clase de artes, en mi caso presenté una pintura que fue bien evaluada, la hice a la rápida y no me tomó mayor trabajo, en cambio, a uno de mis compañeros lo evaluó muy mal, siendo que su pintura era mucho más trabajada, novedosa y había seguido todas las instrucciones. Como no había pauta, no pudo alegar nada".*
- *"En enseñanza media, todas las evaluaciones de historia se realizaban con cuestionarios, el problema radicaba en que las pruebas quedaban en poder del profesor y no se tenía derecho a corrección".*
- *"Un trabajo final de un curso, donde a pesar que mi trabajo fue considerado uno de los mejores del curso en la presentación, fue mal evaluado porque me encontraba en malas relaciones con el profesor".*
- *"Un profesor califica como insatisfactorio un trabajo, pero al comparar notamos que estaba mejor que otros que tenían calificaciones satisfactorias. Son experiencias negativas donde pesa más el criterio del profesor que no es conocido ni entendido por los alumnos".*

Dado que la objetividad de una evaluación consiste en resguardar la ausencia de sesgos o apreciaciones subjetivas en la interpretación de las evidencias y/o procesos que las generaron, se sugieren algunas acciones que un docente puede realizar fácilmente en el aula:

1) *Informar a los alumnos la intencionalidad de la evaluación y los aprendizajes a evaluar*. La doble función de esta acción es que, por una parte, el docente debe hacer explícito el propósito de sus evaluaciones en la planificación, y, por otra, para sus estudiantes no hay ambigüedades en qué se les va a evaluar y con qué fin, por tanto la evaluación deja de ser una "caja negra" y se convierte en una herramienta de aprendizaje con normas claras y bien definidas.

2) *Dar a conocer a los alumnos los criterios de evaluación*. Si entendemos que la evaluación es parte del proceso de enseñanza-aprendizaje, y por ende su fin es ayudar a mejorar los conocimientos de los estudiantes, estos deben saber claramente con qué criterios se les va a evaluar. Preparar una presentación oral que va a ser evaluada en cuanto al uso de las habilidades comunicativas es muy diferente que prepararla para ser evaluado en el dominio conceptual de dicha presentación, por tanto, los énfasis que un estudiante pone y el tipo de aprendizaje que desarrolla estarán en directa relación con la pauta de trabajo que se le haya entregado. Si la pauta y sus criterios son ambiguos o desconocidos, el estudiante realizará lo que cree que debe hacer y si eso no concuerda con lo que el profesor evalúa, la evidencia no será confiable y las conclusiones erróneas.

3) *Elaborar pautas de respuesta o corrección*. Es común elaborar pruebas o situaciones de evaluación, y luego, al momento de corregir se hace un punteo de las ideas que deberían estar presentes en la respuesta o desempeño del estudiante, sin embargo, lo que en una respuesta significa 0.75 puntos cuando se comienza a corregir, puede significar 0.8 cuando se está terminando. El criterio y la exigencia varían producto de diferentes causas (por ejemplo, momento del día, el contexto en el cual se corrige, etc.), por tanto, tener una pauta clara y precisa ayuda a mantener la consistencia en la aplicación y evitar los sesgos propios de un proceso corrección. Además, permite que los estudiantes conozcan esos criterios y puedan contrastarlos con sus resultados, haciendo transparente el proceso de evaluación y permitiéndoles reconocer sus errores y conocer qué se esperaba en cada caso.

4) *Establecer previamente los criterios de asignación de puntajes en función de la relevancia y nivel de complejidad de los aprendizajes*. Este elemento es clave en dos sentidos: primero permite al docente asegurar, previo a la aplicación de la evaluación, que los ítems a evaluar reflejan, en términos de extensión y complejidad, la relevancia que se les asigna con un puntaje determinado en comparación con el puntaje total de la evaluación; y en segunda instancia, los estudiantes toman decisiones en función del peso que tiene cada ítem respecto de la dedicación y orden con la que se enfrentan a ellos. La ausencia de esta asignación genera ambigüedad para los alumnos en los énfasis que da el docente a cada aspecto que se está evaluando.

Conclusiones

La evaluación es parte del proceso de enseñanza y aprendizaje al interior del aula, y como tal puede marcar a un estudiante tanto positiva como negativamente. En esta revisión se ha presentado cómo la ausencia de validez, confiabilidad y objetividad en una situación evaluativa puede generar desmotivación, aprendizajes erróneos y sentimientos de injusticia en los estudiantes, que en nada favorecen un ambiente de aprendizaje que resulte significativo para ellos.

En general, los profesores no calculan la confiabilidad, el error estándar de una medición o los coeficientes de validez y de discriminación de una evaluación. Estas técnicas son propias del desarrollo de pruebas estandarizadas y de aplicación a gran escala, pero tienen una importancia limitada en la evaluación al interior del aula. En el artículo se intenta rescatar los elementos que pueden ser claves y necesarios de resguardar para hacer evaluaciones que aporten al aprendizaje de los estudiantes, como son la validez de contenido, la validez instruccional, la validez consecuencial, la confiabilidad y la objetividad. Como se discutió anteriormente, la validez no se plantea en términos absolutos, sino en grados y depende del propósito para el que fue creada la evaluación,

si sus resultados serán más o menos válidos para sacar conclusiones respecto de ese fin. Tanto la validez como la confiabilidad aumentan o disminuyen según sea la calidad de la evidencia que las sustenta.

La validez de contenido de un instrumento o situación evaluativa se refiere a la cobertura y relevancia de los contenidos y/o habilidades en un área disciplinar específica que han sido seleccionados, y las sugerencias para resguardar este criterio son: elaborar una tabla de especificaciones del instrumento y someter la evaluación a criterio de "jueces".

La validez instruccional hace referencia a la coherencia entre lo que se va a evaluar y lo que los estudiantes han tenido oportunidad real de aprender. Las sugerencias para resguardar este criterio de calidad son: velar por que las situaciones de evaluación contengan los contenidos vistos en las actividades de aprendizaje, sean equivalentes a las actividades realizadas y utilicen un lenguaje conocido por los estudiantes.

La validez consecuencial tiene relación con las consecuencias tanto positivas como negativas que puede tener una evaluación para el aprendizaje de los estudiantes; se sugiere resguardar este criterio definiendo claramente los propósitos y usos de la evaluación e identificando *a priori* las evidencias que darán cuenta de dichos propósitos.

Es pertinente resaltar que la validez instruccional debe ir de la mano de la validez de contenidos, no sería lógico evaluar algo que efectivamente se enseñó, pero que no es relevante para el área disciplinar del curso.

La confiabilidad corresponde a la consistencia y precisión de una evaluación, vale decir, que las evidencias que se recojan del desempeño de los estudiantes sean suficientes para emitir un juicio respecto del nivel de logro de sus aprendizajes. Las acciones que permiten resguardar este criterio son: evaluar en reiteradas ocasiones a un estudiante respecto de un mismo aprendizaje, ya sea en instancias diferentes o en varios ítems dentro de una misma situación evaluativa, velar por la claridad de los ítems, sus instrucciones, el ambiente de aplicación y la precisión en la corrección. Este último elemento también es llamado objetividad y tiene especial importancia al interior de la sala de clases, dadas las consecuencias que puede generar el no resguardarlo. Para ello se sugiere informar previamente a los estudiantes la intencionalidad de la evaluación y los aprendizajes a evaluar, dar a conocer los criterios de evaluación, elaborar pautas de respuesta o corrección y asignar previamente los puntajes en función de la relevancia y nivel de complejidad de los aprendizajes.

Aun cuando la evaluación de aprendizajes al interior del aula es una tarea compleja que requiere tiempo y esfuerzo adicional, es fundamental integrarla como parte del proceso de enseñanza y no verla como una instancia separada. Si se asume esta posición habrá coherencia entre la planificación y la práctica, en la selección de los contenidos y habilidades que se enseñan y que se evalúan, mejorando conductas evaluativas donde escasea la validez, confiabilidad y objetividad.

Bibliografía

Borsboom, D.; Mellenbergh, G.J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review,* 111 (4), 1061-1071.

Brookhart, S. (2003). Developing Measurement Theory for Classroom Assessment. Purposes and Uses. *Educational Measurement: Issues and Practice,* 22 (4), 5-12.

Brown, G. (2004). Teachers' conceptions of assessment: implications for policy and professional development. *Assessment in Education,* 11 (3), 301-318.

Brualdi, A. (1999). Tradicional and Modern Concept of Validity. ERIC/AE Digest. ERIC Clearinghouse on Assessment and Evaluation. Washington DC. ED 435714.

Castillo, S. & Cabrerizo, J. (2007). *Evaluación Educativa y Promoción escolar*. Madrid, España: Pearson Prentice Hall.

Calatayud, M.A. (1999). La creencia en la objetividad de la evaluación: una misión imposible. *Aula Abierta,* 73, 205-222.

Crocker, L.; Llabre, M., & Miller, M.D. (1988). The Generalizability of Content Validity Ratings. *Journal of Educational Measurement* 25 (4), 287-299.

Flórez, R. (1999). *Evaluación pedagógica y Cognición*. Bogotá, Colombia: Mc Graw-Hill.

García, S. (2002). La Validez y la Confiabilidad en la Evaluación del Aprendizaje desde la Perspectiva Hermenéutica. *Revista de Pedagogía,* 23 (67), 297-318.

Gil Pérez, D. & Vilches, A. (2008). ¿Qué deben saber y saber hacer los profesores universitarios? En M.I. Cebreiros y N. Casado (Eds.), *Novos enfoques no ensino universitario* (pp. 25-43). Vigo: Universidad de Vigo.

Gorin, J.S. (2007). Reconsidering Issues in Validity Theory. *Educational Researcher,* 36 (8), 456- 463.

Himmel, E.; Olivares, M.A. & Zabalza, J. (1999). *Hacia una evaluación Educativa. Aprender para evaluar y evaluar para aprender. Volumen I: Conceptos actuales sobre la evaluación del aprendizaje escolar para NB3.* Santiago, Chile: Ministerio de Educación de Chile – Pontificia Universidad Católica de Chile.

Hogan, T. (2004). *Pruebas psicológicas. Una introducción práctica*. México: El Manual Moderno.

Hoyat, F. (1962). *Les Examens*. Institut de l'UNESCO pour l'Education. Paris: Bourrelier.

López, N.; Llopis, R.; Llorens, J.A., Salinas, B. & Soler, J. (1983). Análisis de dos modelos evaluativos referidos a la Química de COU y Selectividad. *Enseñanza de las Ciencias*, 1(1), 21-25.

Luckett, K. & Sutherland, L. (2000). Assessment practices that improve teaching and learning. En S. Makoni (Ed.), *Improving teaching and learning in higher education: a handbook for Southern Africa* (pp.). Johannesburg, South Africa: Witwatersrand University Press.

Lukas, J.F. & Santiago, K. (2004). *Evaluación Educativa*. Madrid, España: Alianza.

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.

Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Moss, P.A. (1997). The role of consequences in validity theory. *Educational Measurement: Issues and Practices,* 17 (2), 6-12.

Moss, P.A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement, Issues and practice,* 22 (4), 13-25.

Moss, P.A. (2007). Reconstructing Validity. *Educational Researcher,* 36 (8); 470-476.

Rosenthal, R. & Jacobson, L. (1968). *Pigmalion in the classroom*. New Jersey: Rinehart and Winston.

Rubie-Davies, C.M. (2006). Teacher Expectations and Student Self-Perceptions: Exploring Relationships. *Psychology in the Schools,* 43 (5), 537-52.

Salinas, D. (2002). *¡Mañana examen! La evaluación entre la teoría y la realidad*. Barcelona: Graó.

Sanmartí, N. (2007). *10 Ideas clave. Evaluar para aprender*. Barcelona: Graó.

Schutz, A., & Moss, P.A. (2004). Reasonable decisions in portfolio assessment: Evaluating complex evidence of teaching, *Education Policy Analysis Archives,* 12 (33). Recuperado el 17 de diciembre de 2008 de http://epaa.asu.edu/epaa/v12n33.

Sireci, S.G. (1998). The Construct of Content Validity. *Social Indicators Research,* 45, 83-117.

Smith, J.K. (2003). Reconsidering Reliability in Classroom Assessment and Grading. *Educational Measurement: Issues and Practice,* 22 (4), 26-33.

Spears, M.G. (1984). Sex bias in science teachers' ratings of work and pupils characteristics. *European Journal of Science Education*, 6, 369-377.

Stiggins, R. J. (2001). *Student-involved classroom assessment*. Upper Saddle River, New Jersey: Prentice-Hall.

Valverde, G. (2000). La interpretación justificada y el uso apropiado de los resultados de las mediciones. En P. Ravela (Ed.), *Los Próximos Pasos: ¿Hacia Dónde y Cómo Avanzar en la Evaluación de Aprendizajes en América Latina?* (pp. 21-30). Lima, Perú: GRADE/Preal.

Fecha de Recepción: 15 de noviembre de 2008 Fecha de Aceptación: 9 de diciembre de 2008