Consideraciones de validez prioritaria para la evaluación[1] formativa y de rendición de cuentas

Priority validity considerations for formative and accountability assessment

Eva L. Baker

Universidad de California en Los Ángeles. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). California, Estados Unidos

Resumen

Este artículo se centra en la validez de los sistemas de evaluación (Baker, 2007a) usados en las escuelas con la intención de mejorar la instrucción, bien a través de evaluaciones formativas o de evaluaciones para la rendición de cuentas. El argumento que se presenta apoya la necesidad de estudios de la sensibilidad instructiva de los tests, necesitando medidas mejores y más equiparadas con la enseñanza en el aula. La agenda de investigación caracteriza a estas preocupaciones como extensas y esenciales. A través de la presentación de los trabajos que se realizan en el CRESST (*Center for Research on Evaluation, Standards and Student Testing*), de la Universidad de California en los Ángeles, su directora, Eva Baker, expone la importancia de la sensibilidad instructiva y los aspectos que se derivan de ella para, posteriormente, pasar a exponer una serie de cuestiones relativas al diseño de pruebas referidas a criterio (CRT). Finalmente, se comenta la situación en la actualidad en los Estados Unidos, donde se realizan evaluaciones con base normativa. Planteadas estas cuestiones, se pone en relación dos importantes aspectos relacionados con la evaluación de los aprendizajes, a saber, la sensibilidad

(1) El trabajo presentado en este artículo ha sido apoyado por el *National Research and Development Centers, PR/Award Number R305A050004,* y administrado por el *Department of Education's Institute of Education Sciences (IES)* de los EE.UU. Los hallazgos y opiniones expresados en este trabajo no reflejan necesariamente las posiciones o políticas del *National Research and Development Centers o del Department of Education's Institute of Education Sciences (IES)* de los EEUU.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

instructiva y las pruebas que persiguen objetivos múltiples. Como conclusión del artículo presentado se presenta un resumen, de modo provisional, sobre las cuestiones que se asumen para la sensibilidad de la instrucción, así como argumentos a favor y en contra de la misma. Se concluye presentando las experiencias del CRESST en el tema.

Palabras clave: evaluación y rendición de cuentas, estudios de validez, evaluación formativa, validez instructiva, evaluación educativa.

Abstract

This paper focuses on validity of testing systems (Baker, 2007a) used in schools intended to improve instruction, either through formative or accountability-based assessments. The argument put forward supports the need for studies of the instructional sensitivity of tests to assure that they are actually measuring school effects. The research agenda suggested by these concerns is extensive and essential. Eva L. Baker presents CRESST (Center for Research on Evaluation, Standards and Student Testing, University of California, Los Angeles) research jobs related to instructional validity and main issues related to Criterion Referenced Test (CRT). Finally, in order to conduct such studies, better and more scalable measures of classroom instruction are needed.

Key Words: assessment and accountability, validity studies, formative assessment, instructional validity, educational assessment.

Introducción

La rendición de cuentas es una función deseada en todo el mundo, y ha adquirido enorme importancia en el área de política educativa. La rendición de cuentas significa hacer responsable a un individuo, grupo, agencia o gobierno de sus acciones y de sus consecuencias conforme intenta alcanzar unas metas específicas. En los Estados Unidos, la rendición de cuentas en educación (Baker, en prensa; Baker *et al,* 2002) se ha centrado fundamentalmente en las escuelas, especialmente en el rendimiento de los estudiantes en ciertos dominios curriculares (matemáticas, lectura y escritura y ciencias) en los niveles de escuela elemental y secundaria. En otros países, la rendición de cuentas puede enfocarse en un rango más amplio de contenidos y sobre el rendimiento individual. En algunos contextos la rendición de cuentas y las consecuencias del rendimiento inadecuado de alumnos o escuelas se distribuyen ampliamente, y las

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

sanciones rodean a estudiantes, profesores, administradores, y responsables políticos. Este artículo se centrará en la experiencia norteamericana y ofrecerá la perspectiva desde la investigación y el desarrollo.

Una revisión del progreso de los estudiantes de Estados Unidos, los más recientes impulsos por la evaluación del rendimiento comenzaron alrededor de 1992, no es halagüeña. Habitualmente, los estudiantes norteamericanos rinden cerca de la media en las comparaciones internacionales (OCDE, 2005; PISA, 2003). Más seria es, por supuesto, la brecha de rendimiento entre estudiantes de minorías y de la mayoría, donde el progreso se ha realizado, pero no ha sido en una tasa suficiente para cambiar el rendimiento general de los grupos en un plazo de tiempo razonable (Baker, Griffin y Choi, 2008). A este respecto, en la búsqueda de la equidad en países con población inmigrante, Estados Unidos no está solo en este afán (OCDE, 2005).

Mientras que *No Child Left Behind* (NCLB) ha sido desarrollada con un plan y una velocidad que, de forma comprensible, ha animado a los Estados a seleccionar exámenes que pueden no ser la mejor medida de la eficacia de las escuelas y del aprendizaje de los estudiantes. Lo que necesita un sistema de rendición de cuentas es tanto credibilidad y validez de las medidas empleadas para calificar el progreso hacia los objetivos, como estándares válidos y alcanzables, y, por supuesto, un plan racional para usar los recursos disponibles, financieros y humanos, para alcanzar los resultados deseados.

La validez es una característica muy conocida vinculada a los tests, coincidiendo en la idea general de su ajuste, cualidad técnica y relevancia de los objetivos. Por supuesto, la validez tiene interpretaciones más complejas, la de mayor alcance es la de Messick (1989) que se centra en la calidad de las decisiones adoptadas a partir de los resultados, más que en las características de la prueba en si misma. Dado que el uso de los tests se ha extendido y sus propósitos nominales se han multiplicado, del mismo modo lo ha hecho la complejidad y la sofisticación requerida de los estudios de validación. Todavía, dado el calendario de desarrollo, se han realizado muy pocos estudios de validez en los Estados Unidos antes de que las evaluaciones se hayan aplicado como medida de la eficacia, y la mayoría de los estudios de validez no abordan el rango completo de propósitos articulados para los exámenes. En lugar de centrarnos en el conjunto entero de propósitos a los que las pruebas de rendición de cuentas (y las evaluaciones formativas) pueden dirigirse, describiremos un conjunto de propósitos que es retador y difícil de abordar. Quizá uniendo experiencias internacionales estemos en condiciones de hacer progresos en estas áreas de validez.

¿Cuáles son los objetivos esenciales de los tests si deben servir como indicadores de alta calidad para la eficacia escolar? En un artículo, publicado hace 15 años,

se enumeran los criterios que las evaluaciones y los tests debían exhibir si querían ser útiles en la escena evaluativa norteamericana (Linn, Baker y Dunbar, 1991; Baker, O'Neil y Linn, 1993). Los criterios relacionan, en parte, los objetivos o estándares que las evaluaciones van a sacar a la luz a raíz de la información recogida en los resultados, ya sea anualmente o en mayores intervalos temporales. Incluimos en esta lista características de complejidad cognitiva: pedir a los estudiantes que realicen procesamientos significativos del contenido y analizarlo utilizando estrategias que requieran múltiples pasos). Una segunda característica relacionada con la riqueza del contenido en que la cognición está sumergida. Este criterio se extendió desde la significación y precisión del contenido como medida. Ambos criterios tuvieron implicaciones sobre cómo el rendimiento tiene que ser juzgado, por ejemplo, si los criterios de puntuación ilustran los niveles de cognición y dominio del contenido deseado. Un tercer criterio fue la equidad, para asegurar que las evaluaciones ofrecían igualdad de oportunidades para estudiantes con distintos contextos socio-culturales o experiencias dentro de los mismos contextos instructivos para demostrar competencia. Inherente a la equidad es evitar la varianza irrelevante para el constructo y el impacto del conocimiento previo, ya sea de los formatos de las pruebas o los contenidos, que sistemáticamente benefician a grupos específicos de estudiantes. La igualdad también se aplicó al acceso y a la capacidad de los estudiantes con necesidades educativas especiales para demostrar competencia en los exámenes. Una interpretación más amplia, implica la extensión de lo que a los estudiantes se les ofrece como razonables oportunidades para aprender el material medido en los test (ver *Raising Standards for American Education, National Council on Education Standards and Testing,* 1992). Una cuarta área para la validez de criterios fue la extensión de lo que el test, o más adecuadamente el sistema de pruebas, aporta a los estudiantes para demostrar transferencia de aprendizajes y generalización de habilidades, estrategias y conocimientos. Estas características son esenciales para asegurar que los estudiantes que han demostrado el uso de procedimientos o resolución de problemas (Vendlinski, Baker y Niemi, 2008) representan una comprensión seria y una expresión en lugar de un producto de procesos superficiales. Mientras que todas las características de la validez son importantes, el resto de este artículo se centra en un conjunto imprescindible si las evaluaciones se usan como base para realizar inferencias sobre la eficacia educativa. El mayor criterio de validez es la «sensibilidad instructiva», componente clave de cualquier razonamiento comprensivo sobre validez para el uso de test en sistemas de rendición de cuentas en educación. La sensibilidad a la instrucción

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

(Baker, 2008) puede razonarse desde la evidencia de que las puntuaciones de un test (examen o evaluación) están afectadas diferencialmente por la instrucción de alta calidad, idealmente impartida en partes significativas del año escolar. El descriptor de alta calidad de la instrucción compleja debe diferenciarse del entrenamiento o de la práctica directa de preguntas muy similares a las de las pruebas que afectan a las puntuaciones de estos tests. Este matiz tautológico ha sido ya analizado en la investigación relacionada con aptitudes o interacciones entre rasgo y tratamiento (ver por ejemplo Berliner y Cahen, 1973; Cronbach y Snow, 1977; y Tobias, 1976. Una conclusión que procede de la investigación en esta área fue que las medidas de resultados requisen especificidad si quieren ser sensibles a los tratamientos).

Este artículo considera por qué la sensibilidad instructiva debe estar documentada, cuáles son las fuentes para las recomendaciones en los diseños de evaluación para incrementar tal sensibilidad y como resultado deseable, unificar las características de las pruebas y de la instrucción. La sensibilidad instructiva aparece cuando se mantienen aspectos de las siguientes situaciones: 1) hay una gran variación en la implementación de un currículo explícito o de un plan instructivo, 2) hay diferencias sustanciales en la preparación y en la calidad resultante de los profesores, 3) se da a los profesores flexibilidad para adaptar el currículo; y 4) no hay un currículo explícito o plan de estudios más allá de la enumeración de los estándares que deben alcanzarse (cierto en muchos de los entornos de Estados Unidos). De la sensibilidad instructiva se presentará, primero, su importancia, después, una breve revisión de las estrategias pasadas para estudiar la sensibilidad instructiva en las escuelas y, por último, una breve consideración de recomendaciones para futuros estudios de validación.

¿Por qué es importante la sensibilidad instructiva?

Cuando las medidas del rendimiento sirven como criterio principal para el juicio sobre la eficacia educativa en los sistemas de rendición de cuentas, se requiere algún nivel de evidencia que muestre una relación defendible entre las puntuaciones de un test y la calidad de la instrucción o de otras experiencias educativas ofrecidas en la escuela. A menos que exista una vinculación causal entre ambas, el uso de tales tests

como medidas dependientes oscila entre lo cuestionable y lo no garantizado. La dirección de la vinculación no es que las pruebas estén diseñadas para adaptarse a la instrucción, sino más bien que los objetivos estimulen actividades instructivas apropiadas y el aprendizaje consecuente sea «muestreado» por el test. Este mapa de objetivos para la instrucción y para la evaluación es lo que se supone debe ser la alineación entre currículo y evaluación (Baker, 2005), pero es difícil de desarrollar completamente por razones que presentaremos posteriormente.

Se esperan las evidencias de validez para una de las finalidades de cada test (American Educational Research Association, American Psycholgical Association y National Council on Measurement in Education, 1999). Por ejemplo, la lógica subyacente de muchos sistemas de rendición de cuentas implica más que la finalidad de identificar la eficacia diferencial de las experiencias instructivas de los estudiantes. En definitiva, hay propósitos relacionados con los sistemas de mejoras a través del tiempo, tal y como se muestra en los patrones de objetivos de crecimiento especificados por *Adequate Yearly Progress* (AYP) en la legislación NCLB o en cualquier otra aproximación que permita el estudio del logro incremental de niveles preestablecidos de rendimiento. Para que tales ciclos de mejora ocurran, los resultados de las pruebas deben ayudar a la selección de materiales instructivos, el diagnóstico y la enseñanza de las actuales cohortes de estudiantes (si el tiempo de los resultados lo permite) o, más fácilmente, de las futuras cohortes de estudiantes (en tanto que el sistema mejora a través del tiempo). Por tanto, es sensible para determinar cómo parte de los estudios de validez, las subordinadas pero necesarias partes de evidencia que muestran si los profesores pueden esbozar inferencias a partir de los resultados de los estudiantes, aplicarlas a los déficit de rendimiento de los estudiantes y desarrollar o aplicar un conjunto de estrategias instructivas alternativas para los estudiantes con déficit, problemas de concepto o inadecuaciones. Estas conductas son esenciales para la teoría de la acción subyacente a todos los sistemas de rendición de cuentas que intentan medir los efectos intencionales de la enseñanza y de la escolaridad (Baker y Linn, 2004). La lógica es también la clave de los conceptos de evaluación formativa (Black y Wiliam, 2003; Pellegrino, Chudowsky y Glaser, 2001). Por supuesto, hay una vinculación intermedia crítica en todas estas finalidades, a saber, que la enseñanza realmente produzca aprendizaje. Déjenos considerar la sensibilidad instructiva y sus derivaciones desde el diseño de los iniciales sistemas de evaluación más íntimamente ligados a la enseñanza y a la mejora. Entonces, como contraste, revisaremos el diseño y los objetivos de los actuales sistemas basados en estándares.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

La derivación de las medidas de sensibilidad instructiva

En los años sesenta y setenta, estimulados por un énfasis en los sistemas auto-instructivos, se produjo un auge de las pruebas referidas a criterio (CRT) o su desarrollo, los tests referidos a dominio (DRT). Estas aproximaciones a la medida fueron tratadas como novedad y apoyadas por importantes académicos, muchos de los cuales habían trabajado en el desarrollo de programas de instrucción o de otros sistemas instructivos integrados antes de entrar en el tema de la medida. Las CRT fueron contrastadas con pruebas referidas a la norma –diseños que intentan una distribución normal de los resultados para permitir ordenar las comparaciones de los examinados (Popham y Husek, 1969)–. La denominación de CRT fue acuñada por Glaser (1963), pero constituida en un trabajo previo (Lumsdaine y Glaser, 1960). El desarrollo de CRT fue diseñada por académicos en matemáticas y áreas científicas vinculadas a un conjunto de teorías (Hively, Patterson y Page, 1968) quienes inventaron y explicaron las propiedades y los beneficios potenciales de las pruebas referidas a dominios.

Cuestiones en el diseño de Pruebas Referidas a Criterio (CRT)

Entre las principales propiedades de las pruebas referidas a criterio está el punto de partida del diseño: un dominio bien especificado de contenido y habilidades intelectuales (ahora orientación cognitiva). Proveniente desde una perspectiva conductista del aprendizaje, el diseño de esos tests estaba muy especificado, y a lo largo del tiempo cambió de lugar desde la simple enumeración detallada hasta incluir un mayor desarrollo y racionalización de las demandas cognitivas. Actualmente, el nivel de especificidad de esas medidas ha hecho más que necesario crear secuencias instructivas realizables y «copiables» en el tiempo. En lugar de empezar con un constructo general y especificaciones del test que distribuía los ítems en uno o más formatos a través de una amplia gama de temas, la nueva alternativa requiere contenidos bien definidos y habilidades delimitadas incluyendo dominios de contenido, problemas tipo, formatos y reglas de puntación que delimiten el grupo de tareas o ítems. Así como surgieron las ideas, lo hizo la noción probabilística del conjunto de ítems (por ejemplo los subconjuntos difusos), que caracteriza la precisión del conjunto de tareas o la prueba de ítems dentro de las características explícitas del dominio. Se espera que una muestra de tareas

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

construida de manera extensa y cuidadosa debería medir profunda y directamente tanto dentro del dominio total como de sus diferentes subpartes.

Estos límites del domino o reglas pretendían también tener un poder «instruccional» porque podían ser transparentes (lenguaje claro y apoyado en ejemplos) y compartidos directamente con los profesores que esperaban controlar la instrucción. Los profesores que desarrollaron o aplicaron una instrucción en la que se exhibían esas características de los dominios explicados esperaban producir una ganancia previsible en el rendimiento en las CRT, al igual que los sistemas de instrucción que se esperaba que fuesen efectivos después de un ciclo o dos de revisión. Dado que el impuso de las CRT vino del diseño de sistemas de instrucción sistemática, se esperaba que los alumnos con menor rendimiento satisfactorio debían conseguir cambios adicionales, con una instrucción más refinada, aprender y de esa forma demostrar sus habilidades con un rendimiento mejorado en los tests. Esta frecuente visión ha caracterizado la formación durante un buen número de años. Esta idea de definir el universo a partir de que las tareas de los tests pueden ser bastante esbozados (Hively et al., 1968) tuvo la mayor parte del poder en la idea de las Pruebas Referidas a Criterio.

Los informes CRT

Otra característica más destacada y fácil de implementar de las CRT, se centró en los informes. Los resultados de las CRT deberían informar de manera diferente a los tests comerciales disponibles. Las CRT intentaban dar a conocer algunas ideas de las competencias o dominio respecto al (tal vez arbitrario) cuerpo de conocimientos y habilidades en lugar de apuntar a un particular punto dentro de un amplio constructo de capacidad. De esa manera, los informes de las CPR incluyen, al comienzo, medias en bruto como el porcentaje correcto o el porcentaje de logro de algunos objetivos en gran parte arbitrarios (el 90% logrará 90%). Esos números parecen dar una imagen más definitiva de la adquisición de las competencias proyectadas del alumno cuando se contrastan con valores normativos transformados que reflejan el lugar del sujeto en la distribución de los examinados, pero eran fácilmente objeto de manipulaciones. La fijación de puntaciones de corte con el fin de dividir el grupo en varios niveles de competencia, por si mismo generó una industria artesanal sobre cómo desarrollar puntuaciones de corte y descripciones verbales adicionales que caracterizasen los

resultados de los alumnos. Es de suponer que esos niveles tuvieron que ser validados también. Fue la parte de presentación de informes de CRT la que capturó la atención de desarrolladores y usuarios de los tests. Es relativamente fácil convertir cualquier registro de rendimiento en distribuciones de frecuencia dentro de un particular rango o rangos de puntuaciones, establecidos por un standard de rendimiento o puntuaciones de corte. De hecho, los protocolos de los informes están repletos de requisitos de diseño, que nos sitúan para las prácticas de evaluación a gran escala.

La situación actual. Evaluaciones con base normativa

Los sistemas de pruebas con base normativa a menudo son entendidos y representados como si fuesen CRT, y es seguramente la impresión dada por la noción de base normativa, por lo menos como se practica en gran parte de los EE.UU. Estos sistemas, de hecho, usan estándares como el origen de su diseño, cuyas descripciones verbales circunscriben únicamente en términos generales las tareas dentro de los ítems. Antes que utilizar las reglas de diseño para el conjunto de ítems, los tests de base normativa han adoptado las características de información de las CRT. Evidentemente, usan criterios de corte entre categorías, como por debajo de básico (*Below Basic*), etc. Estos son fijados por complejos planteamientos relativos a los ítems cuyo propósito podría o no perdurar en procedimientos de escala posteriores. En el diseño frontal, los detalles de la evaluación son menos implícitos. Rara vez apelan a un constructo real validado, consideración que podría ser apropiado a la estructura de su diseño.

La instrucción alineada a estándares y evaluaciones bajo esas condiciones presenta un problema desafiante. Los estándares se piensa que deberían guiar la instrucción y debería ajustarse al centro de la clase. Los tests deberían ser meros indicadores de la adquisición y aplicación de habilidades y conocimientos.

Se pueden encontrar un conjunto de procedimientos (Herman & Webb, 2007) para decidir cuánto ajuste se ha logrado (Baker, 2005). En su mayor parte, estos enfoques se dirigen a la referencia de contenidos que sirven de modelo teórico a ítems relevantes en el test. Aunque se han hecho algunos esfuerzos por establecer el alcance de la medición (Webb, 1999), no es en absoluto un criterio universal y, por tanto, muchas pruebas estatales pueden tener un número relativamente reducido de ítems, o tan pocos que uno o una fracción de uno, localiza un estándar o uno de sus elementos.

La razón para decidir usar estas fuentes de medida incluye el tiempo asignado para la realización de pruebas, la velocidad de ejecución y los costes.

Sin embargo, cuando los estándares no son medidos de forma adecuada (tal vez debido a obstáculos como su número, el tiempo o los costes),¿qué debería hacer un profesor frente a las sanciones? Es mucho más prudente para ese profesor evitar sanciones mediante la enseñanza de los temas que aparecen en pruebas con cierta frecuencia. Ellos pueden obtener esta información de las pruebas de la inspección, personalmente o por medio de diseños de inspección sustitutos que determinen qué se incluye en realidad en los tests. El resultado es conocido, en la práctica el contenido de la prueba se ha convertido en la norma en muchas escuelas y, en particular, en aquellas con riesgo de sanción. Las consecuencias de este planteamiento son múltiples: los esfuerzos para enseñar estándares se están perdiendo a favor de la enseñanza de contenidos testados; la coherencia y el conocimiento acumulado se pierde; los estudiantes con menor rendimiento tiene poca atención sistemática a subtareas de aprendizaje que podrían sustentar varios estándares y resultados futuros. Las adquisiciones de procedimientos o de trucos en el procedimiento puede ser más fácil de enseñar que conceptos dificultosos en abundantes aplicaciones.

Esto no tiene un gran valor en la esfera política, donde muchos creen que los detalles de test específicos no importan tanto como sus títulos, y los matices, como la profundidad de muestreo, puede ser concebido como una característica técnica pero no esencial. Cuando las medidas están fuertemente correlacionadas, se ha opinado públicamente por los políticos, que son intercambiables unas por otras. Así que, incluso en las evaluaciones con base normativa, muchos de los enfoques tradicionales hacia la tarea o el desarrollo de ítems y el muestreo se utilizan bastante despreocupadamente.

Objetivos múltiples y sensibilidad instructiva

Al igual que ha sucedido con los requisitos para las medidas de rendición de cuentas desarrolladas, ha llegado a ser cada vez más patente que se esperaba que los tests sirviesen para diferentes propósitos, algunos de ellos señalados anteriormente. La mayoría de los expertos en psicometría señalan que cada uno de los objetivos del test es digno de su propio test, con la finalidad de optimizar la calidad de la consiguiente

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

decisión que debe llevarse a cabo, ya se trate de la entrada, la colocación, o el sistema de monitorización. Parece que las pruebas creadas para un propósito rara vez pueden ser adaptadas para servir a un propósito diferente. Por supuesto, el proceso de adaptación era la única forma en la que las pruebas de rendición de cuentas se podían crear de manera rápida y económica. Sin embargo, si tales pruebas estaban principalmente diseñadas usando especificaciones generales más que dominios bien especificados u ontológicos, ¿cómo puede ser abordado el concepto de sensibilidad de la instrucción?

¿Instrucción de alta calidad?

La cuestión es sencilla, ¿cómo saber que las evaluaciones de medida nominales ansían resultados que actualmente son sensibles a una instrucción de alta calidad? Si hemos esbozado las características de las medidas que podrían hacerlas sensibles a la instrucción, ¿cómo debería ser caracterizada la calidad de la instrucción? En primer lugar, y como de costumbre, la instrucción debería asociarse sustantivamente a los objetivos a los que se dirige. Debería representar la clave cognitiva y los elementos de contenido dispuestos en secuencias diseñadas para garantizar condiciones previas. En el mejor de los casos, la instrucción debe proporcionar los principios básicos y las estrategias de apoyo pertinentes y los prerrequisitos de conocimiento, de forma que los estudiantes entiendan por qué están utilizando enfoques particulares y donde encajan en el dominio o en una secuencia de la instrucción. En la medida de lo posible, la instrucción debe guiarse por los resultados de la investigación sobre el aprendizaje que traten de minimizar el peso memorístico o cognitivo (Sweller, 1999), para apoyar el diseño desarrollado (Chi, Glaser, & Farr, 1988), y proporcionar información pertinente. Para apoyar la transferencia, la instrucción debería llevar consigo una amplia gama de formatos para los tests, así como otras estrategias aplicadas e integradas para mostrar los logros.

Éstos incluyen retroalimentación adjunta al dominio práctico adecuado, la comprensión de las intenciones (motivación), la diferenciación individual, y una secuencia gradual de aprendizaje todos implicados en la participación activa de los estudiantes (véase Popham y Baker, 1970, para un tratamiento precoz de estos conceptos en un contexto de preparación del profesor). Estas aplicaciones de los principios pedagógicos son viables y razonablemente eficaces.A pesar de esto, el grado de intensidad,

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

mezcla, y frecuencia de estos y otros principios de la instrucción, andamiajes, y niveles de engranaje, por ejemplo, podría sugerir refinados estudios de validez. Si podemos ponernos de acuerdo sobre la parte de «calidad» de la instrucción, podemos invertir nuestros análisis y ver qué tipo de prueba responde mejor.

Condiciones asumidas para la Sensibilidad de la Instrucción: un resumen provisional

Para estudiar la sensibilidad de la instrucción, hay algunas cuestiones recurrentes que deben incluirse y deberían responderse afirmativamente, de manera que al menos haya un alto tratamiento de calidad. Estas preguntas son:

- ¿Los resultados de los tests comunican dominios operacionales para la acción instructiva? ¿Son entendidos por los profesores?
- ¿Hay experiencia y tiempo para la enseñanza basada en los resultados?
- ¿Para la atención a las diferencias?
- ¿Pueden los profesores investigar dónde se necesita ayuda (evaluación formativa de alta calidad)?

¿Cómo podemos saber? Evidencias a favor y en contra de la Sensibilidad de la Instrucción.

Las indicaciones anteriores eran una lista de criterios de validez para las evaluaciones (Baker, O'Neil, y Linn, 1993). El criterio de validez especificado en este artículo proporciona gran parte del pegamento intelectual de la agenda de I+D del CRESST para las próximas dos décadas. Las cuestiones claves que podrían ser respondidas por los datos de la inspección estatal, con certeza, nos da una estimación bruta de la probabilidad de encontrar la sensibilidad de la instrucción, son las siguientes:

- ¿Cuál es la relación entre la cantidad de tests prácticos y el tiempo total de instrucción funcional y cómo se conecta con los resultados de las pruebas?
- ¿Hay diferencias en las tasas de exposición a esta práctica por subgrupos?

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

- ¿Existen tratamientos de instrucción focalizados en la rápida adquisición de prerrequisitos?
- ¿Es evaluado el rendimiento en la transferencia de tareas?
- ¿Existe evidencia en los datos de que el aprendizaje individual es acumulativo y sostenido?

Los niveles actuales de evidencia o bien están vacíos o sólo compete modestamente a la mayoría de nuestros estados.

Las experiencias del CRESST en cuestiones de Sensibilidad Instructiva

Como parte de nuestra investigación, prometimos dirigir la atención hacia los criterios de validez que más tarde aparecieron en el artículo de Linn et al., y los criterios utilizados como cimiento para vincular los estudios de evaluación del diseño y la validación. CRESST emprendió el desarrollo de una nueva encarnación de CRT en 1990, llamado *Model-Based Assessments* (Baker, 1997, 2007b) anteriormente denominado «Evaluación Cognitiva-Sensitiva». Este fue nuestro intento de desarrollar evaluaciones sensibles a la instrucción, es decir, para reflejar un dominio complejo cuyos rasgos se podrían utilizar para orientar la planificación de la instrucción de los profesores o de otros. Además, hemos tenido preguntas acerca de qué atributos de las evaluaciones podrían interpretarse como un dominio independiente, es decir, que fuese útil a través de los diferentes temas y materias, los cuales se centrasen en criterios extraídos de la literatura sobre el aprendizaje de un determinado dominio *What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001), y si el desarrollo de esas medidas podría centrarse en primer lugar en demandas cognitivas que podrían ser dominios independientes y, por tanto, reducir el costo de elaboración de subsiguientes medidas de alta calidad.

La medición de la práctica de las aulas

El CRESST, junto con un gran número de investigadores de la comunidad educativa, se ha interesado en la medición de la práctica en las aulas, o en la oportunidad de aprender

(Aguirre-Muñoz *et al,* 2006; Herman & Abedi, 2004; Herman & Klein, 1997; Yoon & Resnick, 1998). Los enfoques utilizados incluyen las observaciones, los autoinformes de conocimientos y pedagógicos (Baker *et al,*1996), las observaciones (Boscardin *et al,* 2004), las tareas del estudiante y su trabajo (Aschbacher, 1999;Aschbacher & Clare, 2001; Matsumura *et al,* 2002; Matsumura y Pascal, 2003; Matsumura *et al,* 2006), y la interpretación por parte de los maestros de los resultados y los planes de instrucción (Herman & Baker, 2003).

Estos estudios se llevaron a cabo en el CRESST con el fin de determinar en qué medida nuestras evaluaciones aumentaban, o en otras palabras, el grado en que la fidelidad entraba en conflicto con la realidad de la clase (Baker *et al,* 1996). También hemos llevado a cabo estudios de proceso-producto utilizando medidas externas disponibles (Goldschmidt *et al,* 2007) (en el rendimiento de los tests del CRESST y los exámenes de graduación de las escuelas de secundaria), a menudo en el contexto de evaluar una intervención. En estos estudios buscábamos las relaciones entre los autoinformes, la observación, o los artefactos (Matsumura *et al,* 2002; Stecher *et al,* 2007). Todos estos estudios, si bien interesan en cuanto que hay pocas relaciones, casi nunca producen datos convincentes debido a la confusión de la calidad del profesorado y logro del alumno, aspecto señalado por el trabajo de la «razón de ventaja» (*odds-ratio*) de Rogosa (Rogosa, 1999a, b, c).

Referencias bibliográficas

American Educational Research Association (2008). *Unleashing the power of formative assessment: A strategy for integrating cognitive research,* assessment, and instruction. Session 13.028 at the annual meeting of the American Educational Research Association, New York.

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

Aguirre-Muñoz, Z., Boscardin, C. K., Jones, B., Park, J. E., Chinen, M., Shin, H. S., Lee, J., Amabisca, A. A. y & Benner, A. (2006). *Opportunity to learn measures* (CSE Rep. 678). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Aschbacher, P. R. (1999). *Developing indicators of classroom practice to monitor and support school reform* (CSE Rep. 513). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. L. (in press). Learning and assessment in an accountability context. In K. Ryan & L.A. Shepard (Eds.), *The future of test-based educational accountability.* Mahwah, NJ: Erlbaum.

— (1997). Model-based performance assessment. *Theory Into Practice,* 36, 247-254.

— (2005). *Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform.* In C.A. Dwyer (Ed.), Measurement and research in the accountability era (pp. 315-335). Mahwah, NJ: Erlbaum.

— (2007). Model-based assessments to support learning and accountability: The evolution of CRESST's research on multiple-purpose measures. *Educational Assessment* (Special Issue), 12(3&4), 179-194.

— (2008). Empirically determining the instructional sensitivity of an accountability test. Paper presented at the annual meeting of the American Educational Research Association. In W. James Popham, *Empirically Determining the Instructional Sensitivity of an Accountability Test: Alternative Approaches.* New York.

Baker, E. L. (2008). Empirically determining the instructional sensitivity of an accountability test. Artículo presentado en la reunión anual de la American Educational Research Association. In W. James Popham session 28.072, *Empirically Determining the Instructional Sensitivity of an Accountability Test: Alternative Approaches.* New York.

Baker, E., Goldschmidt, P., Martínez, F. & Swigert, S. (February, 2002). *In search of school quality and accountability: Moving beyond the California Academic Performance Index (API)* (Deliverable to OERI, Contract No. R305B6002). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Baker, E. L., Griffin, N. C. & Choi, K. (in press). *The achievement gap in California: Context, status, and approaches for improvement.* Paper prepared for the California Department of Education, P-16 Closing the Gap Research Council "Connecting the Dots and Closing the Gap." Davis, CA: University of California, Center for Applied Policy in Education (CAP-Ed).

Baker, E. L. & Linn, R. L. (2004). Validity issues for accountability systems. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 47-72). New York: Teachers College Press.

Baker, E. L., Niemi, D., Herl, H., Aguirre-Muñoz, Z., Staley, L., Linn, R. L. & Rogosa, D. (1996). *Report on the content area performance assessments* (CAPA): A collaboration

among the Hawaii Department of Education, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the teachers and children of Hawaii (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. L., O'Neil, H. F., J. R. & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist,* 48, 1210-1218.

Baker, E. L., Phelan, J., Choi, K., Niemi, D., Vendlinski, T., Griffin, N., Herman, J. L. & Howard, K. (in progress). *Design and validation of POWERSOURCE© assessments and instructional materials.* Los Angeles, University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Berliner, D. C. & Cahen, L. S. (1973). Trait-treatment interaction and learning. *Review of Research in Education,* 1, 58-94.

Black, P. & Wiliam, D. (2003). In praise of educational research: formative assessment. *British Educational Research Journal,* 29, 623-637.

Boscardin, C. K., Aguirre-Muñoz, Z., Chinen, M., Leon, S. & Shin, H. S. (2004). *Consequences and validity of performance assessment for English learners: Assessing opportunity to learn (OTL) in grad 6 language arts* (CSE Rep. 635). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Chi, M.T. H., Glaser, R. & Farr, M. (Eds.). (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum.

Clare, L. & Aschbacher, P. R. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment,* 7, 39-59.

Cronbach, L. J. & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-521.

Goldschmidt, P., Martinez, J. F., Niemi, D. & Baker, E. L. (2007). Relationships among measures as empirical evidence of validity: Incorporating multiple indicators of achievement and school context. *Educational Assessment* (Special Issue), 12 (3&4), 239-266.

Herman, J. L. & Abedi, J. (2004). *Issues in assessing English language learners' opportunity to learn mathematics* (CSE Rep. 633). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Herman, J. L. & Baker, E. L. (2003). *The Los Angeles Annenberg Metropolitan Project: Evaluation findings* (CSE Tech. Rep. No. 591). Los Angeles: University of California,

National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L. & Klein, D. C. D. (1997). *Assessing opportunity to learn: A California example* (CSE Rep. 453). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Herman, J. L. & Webb, N. M. (2007). Alignment methodologies. *Applied Measurement in Education,* 20, 1-5.

Hively, W., Patterson, H. L. & Page, S. H. (1968). A «universe-defined» system of arithmetic achievement tests. *Journal of Educational Measurement,* 5, 275-290.

Linn, R. L., Baker, E. L. & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher,* 20, 15-21. (ERIC Document Reproduction Service No. EJ 436 999)

Lumsdaine, A. A. & Glaser, R. (Eds.). (1960). *Teaching machines and programmed learning: A source book.* Washington, DC: National Education Association of the United States.

Matsumura, L. C., Garnier, H. E., Pascal, J. & Valdes, R. (2002). *Measuring instructional quality in accountability systems: Classroom assignments and student achievement* (CSE Rep. 582). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Matsumura, L. C. & Pascal, J. (2003). *Teachers' assignments and student work: Opening a window on classroom practice* (CSE Rep. 602). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Matsumura, L. C., Slater, S. C., Wolf, M. K., Crosson, A., Levison, A., Peterson, M., Resnick, L. & Junker, B. W. (2006). *Using the instructional quality assessment toolkit to investigate the quality of reading comprehension assignments and student work* (CSE Rep. 669). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: MacMillan.

National Council on Education Standards and Testing. (1992). *Raising standards for American education.* Washington, DC: U.S. Government Printing Office. (ERIC Document Reproduction Service No. ED338721)

Niemi, D., Wang, J., Steinberg, D. H., Baker, E. L. & Wang, H. (2007). Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment,* 12 (3&4), 215-237.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425 (2002).

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

Popham, W. J. & Baker, E. L. (1970). *Systematic instruction.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement,* 6, 1-9.

Rogosa, D. (1999a). *Accuracy of individual scores expressed in percentile ranks: Classical test theory calculations* (CSE Tech. Rep. No. 509). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

— (1999b). *Accuracy of Year-1, Year-2 comparisons using individual percentile rank scores: Classical test theory calculations* (CSE Tech. Rep. No. 510). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

— (1999c). *How accurate are the STAR national percentile rank scores for individual students? An interpretive guide.* Palo Alto, CA: Stanford University.

Stecher, B., Borko, H., Kuffner, K. L., Martinez, F., Arnold, S. C., Barnes, D., Creighton, L. & Gilbert, M. L. (2007). *Using artifacts to describe instruction: Lessons learned from studying reform-oriented instruction in middle school mathematics and science* (CSE Rep. 705). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Sweller, J. (1999). *Instructional design in technical areas.* Camberwell, Australia: ACER Press.

Tobias, S. (1976). Achievement treatment interactions. *Review of Educational Research,* 46, 61-74.

Vendlinski, T. P., Baker, E. L. & Niemi, D. (2008). Templates and objects in authoring problem-solving assessments. In E. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 309-333). New York: Erlbaum.

Vendlinski, T. & Stevens, R. (2000). The use of artificial neural nets (ANN) to help evaluate student problem solving strategies. In B. Fishman & S. O'Conner-Divelbiss (Eds.), *Proceedings of the fourth international conference of the learning sciences* (pp. 108-114.). Mahwah, NJ: Erlbaum.

Webb, N. L. (1999). *Research Monograph No. 18: Alignment of science and mathematics standards and assessments in four states.* Madison, WI: National Institute for Science Education.

Yoon, B. & Resnick, L. (1998). *Instructional validity, opportunity to learn and equity: New standards examinations for the California mathematics renaissance* (CSE Rep. 484). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Fuentes electrónicas

Baker, E. L. (2007a, August/September). The end(s) of testing (2007 AERA Presidential Address). Educational Researcher, 36 (6), 309-317. Retrieved October 2, 2007, de http://www.aera.net/uploadedFiles/Publications/Journals/Educational_Researcher/3606/09edr07_309-317.pdf

OECD. (2005). *PISA 2003 data analysis manual: SAS® users.* Paris: Author. Available online at http://www.pisa.oecd.org/dataoecd/53/22/35014883.pdf

Dirección de contacto: Eva L. Baker. Universidad de California en Los Ángeles. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Los Ángeles California, Estados Unidos. E-mail: eva@ucla.edu