PSYKHE
2016, 25(1), 1-14
doi:10.7764/psykhe.25.1.681

ISSN 0717-0297
www.psykhe.cl

Adaptación y Pilotaje de un Portafolio Para Evaluar Prácticas de Evaluación de Aprendizajes en el Aula en Profesores de Matemática de Segundo Ciclo Básico

Adaptation and Piloting of a Portfolio for Evaluating Classroom Assessment Practices of Middle School Mathematics Teachers

María Asunción Pérez-Cotapos y Sandy Taut
Pontificia Universidad Católica de Chile

Entre las habilidades más importantes y valiosas de los profesores se encuentra su capacidad para realizar prácticas de calidad de evaluaciones de aprendizaje en el aula (PEAA). El objetivo del estudio fue la adaptación y pilotaje de un portafolio que capture las PEAA de profesores de matemática de segundo ciclo básico. El pilotaje se realizó con una muestra intencionada de 17 profesores de diversos contextos socioeconómicos de Santiago, Chile. Los profesores recolectaron 2 portafolios con evidencia de sus PEAA y contestaron 2 cuestionarios de autopercepción respecto de ellas, además de una entrevista. Los portafolios fueron evaluados por 2 profesores de matemática entrenados con una rúbrica de evaluación. Se hicieron análisis de generalizabilidad, correlaciones, comparación de medias y análisis de contenido de las entrevistas. Los resultados muestran baja correlación entre el auto-reporte y los resultados del portafolio. También se encuentran indicios iniciales cualitativos de un posible efecto positivo del portafolio en las prácticas de reflexión y la auto-percepción. Se observa la necesidad de perfeccionar el instrumento, la rúbrica, la selección y capacitación de los correctores para aumentar la confiabilidad del portafolio.

Palabras clave: portafolio, evaluación de aprendizaje, matemática, prácticas docente, generalizabilidad

One of the most important and valuable skills of a teacher is the capacity to use high quality classroom assessment practices. The present study adapted and piloted a portfolio that captures teachers' classroom assessment practices in middle school mathematics. The purposive sample comprised 17 middle school teachers from schools of diverse socioeconomic status in Santiago, Chile. Participants completed 2 portfolios with evidence of their classroom assessment practices, answered 2 self-perception questionnaires about their practices, and were interviewed. The portfolios were scored by 2 mathematics teachers who were trained to apply a scoring rubric. The analyses included a generalizability study, correlations, mean comparisons, and content analysis of the interviews. Results indicate a low correlation between teachers' self-reported practices and the portfolio evidence. In addition, initial qualitative results suggest a positive effect of the portfolio on participants' capacity to reflect about their assessment practices. It is necessary to further improve the instrument, the scoring rubric, and rater selection and training in order to increase the portfolio's reliability.

Keywords: portfolio, assessment, mathematics, teaching practices, generalizability

Una de las habilidades más importantes y valiosas de los profesores es su capacidad para realizar evaluaciones de aprendizaje de calidad. Las prácticas de evaluación de aprendizaje en el aula (en adelante PEAA) son la medición educacional más influyente en el aprendizaje de los alumnos, ya que entregan información de calidad sobre el proceso de aprendizaje de cada alumno, sus fortalezas y debilidades (Campbell, 2013) y retroalimentación para mejorar las prácticas docentes (Pecheone & Chung, 2006; Wolf & Taylor, 2008).

Si bien existe poca investigación en cuanto a la calidad y efectividad de las PEAA, muchos autores acentúan la importancia de ellas, debido a las numerosas repercusiones que tienen en los procesos de aprendizaje (Black & Wiliam, 1998; McMillan, 2013), haciendo un llamado a aumentar las investigaciones específicas en PEAA como requisito indispensable para la mejora de las prácticas pedagógicas y un consecuente aumento del

María Asunción Pérez-Cotapos Valenzuela, Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago, Chile; Sandy Taut, Centro de Medición MIDE UC, Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago, Chile.

Las autoras agradecen el financiamiento parcial de este trabajo por parte del Fondo Nacional de Desarrollo Científico y Tecnológico de Chile (Proyecto FONDECYT Nº 1120441).

La correspondencia relativa a este artículo debe ser dirigida a María Asunción Pérez-Cotapos o a Sandy Taut, Escuela de Psicología, Pontificia Universidad Católica de Chile, Avda. Vicuña Mackenna 4860, Macul, Santiago, Chile. E-mail: msperezc@uc.cl y staut@uc.cl, respectivamente.

aprendizaje y motivación de los alumnos (McMillan, 2013). Randel y Clark (2013) plantean que las PEAA de calidad incluyen claros y apropiados objetivos de aprendizaje, retroalimentación descriptiva a los estudiantes (Hattie & Timperley, 2007; Wiliam, Lee, Harrison & Black, 2004), involucramiento de los estudiantes en su propia evaluación de aprendizaje (Black & Wiliam, 1998; Stiggins, 2001) y un sistema de puntuación y comunicación informativo.

Debido a la complejidad y centralidad que tienen las PEAA en el proceso de enseñanza-aprendizaje, se ha estimado que en la práctica pedagógica cotidiana los profesores dedican más del 50% del tiempo en actividades relacionadas con ellas (Stiggins, 1991). A pesar de esto, muchos profesores reportan no sentirse bien preparados para esta tarea (Campbell, 2013; Montecinos, Rittershaussen, Solís, Contreras & Contreras, 2010). Además, la investigación ha mostrado que el auto-reporte de los docentes respecto de sus prácticas evaluativas no correlaciona con la evaluación de estas prácticas en base de la evidencia empírica directa, recolectada en un portafolio (Randel & Clark, 2013; Romo, Treviño & Villalobos, 2014).

Independientemente de la auto-percepción de los profesores, los resultados de la Evaluación Docente 2011 en Chile (Ministerio de Educación [MINEDUC], 2012) muestran que dos de las dimensiones con menor desempeño son *Calidad de la evaluación de la unidad* y *Reflexión a partir de los resultados de la evaluación*. Incluso docentes con desempeño final competente y destacado tienen, en promedio, un nivel básico e insatisfactorio en estas dimensiones. En esta misma línea, un estudio realizado con estudiantes de pedagogía chilenos en su último año de estudio observó que más de un 75% de los estudiantes no alcanzó el nivel competente para el diseño de un plan de evaluación, el análisis de los resultados de las evaluaciones y la toma de decisiones respecto de ellos, quedando un 33% en un nivel insatisfactorio (Montecinos et al., 2010). Este hallazgo es de gran relevancia, ya que un estudio realizado por Santelices y Taut (2011) encontró que las prácticas evaluativas de los docentes chilenos son predictores relevantes para distinguir entre profesores de nivel destacado e insatisfactorio. A su vez, Romo et al. (2014) muestran cómo los docentes con menores desempeños se agrupan mayoritariamente en establecimientos con alumnos más desventajados. Es por esto que se hace indispensable contar con más estudios que aporten al conocimiento sobre las PEAA, las cuales destacan como una práctica pedagógica específica que diferencia docentes de alto y bajo desempeño, para mejorar la calidad de las prácticas pedagógicas en los sectores donde más se necesita.

Para evaluar las PEAA se pueden utilizar distintos instrumentos, incluyendo observaciones, entrevistas, cuestionarios y pruebas de contenido (Randel & Clark, 2013). Sin embargo, en la literatura internacional se considera el portafolio como el más adecuado (Martínez, Borko, Stecher, Luskin & Kloser, 2012; Randel & Clark, 2013). Esta herramienta permite acceder a información relevante del proceso de evaluación, como son las planificaciones de las evaluaciones, la intencionalidad evaluativa, los materiales de evaluación diagnóstica, formativa y sumativa y las reflexiones y decisiones instruccionales (Tucker, Stronge, Gareis & Beers, 2003).

Una segunda razón por la cual el portafolio resulta una herramienta interesante es su capacidad para reflejar la complejidad de los procesos evaluativos, que favorece la reflexión de los docentes en torno a sus prácticas (Tucker et al., 2003) y el desarrollo de mejores prácticas evaluativas (Pecheone & Chung, 2006; Wolf & Taylor, 2008). Así, es esperable que la participación de profesores en la aplicación de un portafolio tenga un impacto positivo en sus prácticas de reflexión pedagógica y en la auto-percepción de sus habilidades para realizar PEAA.

Es preciso constatar que el portafolio es una herramienta validada para evaluar prácticas docentes generales y ha sido estudiado en diversas investigaciones (Kennedy, 2010; Martínez, 2013). Por ejemplo, es utilizado en evaluaciones de gran escala, como el *Performance Assessment for California Teachers* (PACT; Pecheone & Chung, 2006), la acreditación de excelencia por el *National Board for Professional Teaching Standards* (Bond, Smith, Baker & Hattie, 2000; Gitomer, 2008; Goldhaber & Anthony, 2004; Hattie, 2008; Hattie & Clinton, 2008; Jaeger, 1998; National Research Council, 2008; Pool, Ellett, Schiavone & Carey-Lewis, 2001) y la Evaluación Docente en Chile (Manzi, González & Sun, 2011; Santiago, Benavides, Danielson, Goe & Nusche, 2013).

Hasta la fecha existen pocos instrumentos para evaluar PEAA a través de evidencia física, dos de los cuales son el *Teacher Assessment Work Samples* (Randel et al., 2011) y el *QAS Notebook* (Martínez et al., 2012). El *QAS Notebook* es un portafolio docente, en el cual los profesores deben recolectar durante 10 clases (módulos) evidencia física de las evaluaciones que realizan en estas. La concepción de las PEAA que utilizan es amplia e incluye tanto pruebas formales como interacciones entre el profesor y el alumno que sirvan de insumo al profesor para conocer el nivel de aprendizaje de sus alumnos. Cada portafolio es corregido a través

de la aplicación de una rúbrica de corrección con nueve dimensiones específicas y una evaluación global. Los estudios de validez del *QAS Notebook* se realizaron a través de dos estudios de generalizabilidad, análisis factorial y correlaciones con otras variables (Martínez et al., 2012).

El presente estudio tuvo como objetivo central la adaptación y pilotaje del *QAS Notebook* al contexto chileno. La motivación fue avanzar en la búsqueda de nuevas herramientas válidas para evaluar y retroalimentar las PEAA en Chile. Las siguientes preguntas de investigación guiaron el trabajo:

1. ¿Puede un portafolio adaptado al contexto chileno en base del *QAS Notebook* medir las PEAA de los docentes de matemática de manera confiable?
2. ¿Cómo se relaciona el auto-reporte de los docentes respecto de sus prácticas evaluativas con la evaluación de estas prácticas en base de la evidencia recolectada en el portafolio?
3. ¿Existen cambios en la auto-percepción de las habilidades evaluativas de los docentes después de haber implementado el portafolio?

Método

Participantes

Profesores. El muestreo se realizó por conveniencia, durante Agosto y Septiembre de 2012, asegurando la participación de profesores de colegios de distintos contextos socioeconómicos. Participaron 17 profesores de matemática de Santiago, Chile, de segundo ciclo básico (6º, 7º y 8º grado de la enseñanza básica) de seis establecimientos municipales (con financiamiento y gestión estatal), ocho particulares subvencionados (con financiamiento mixto y gestión privada) y tres particulares pagados (con financiamiento y gestión privada).

Por su participación en el estudio los profesores recibieron una retribución de 20.000 CLP (US$ 34), por sus gastos de transporte y fotocopias, y un certificado de capacitación de la Pontificia Universidad Católica de Chile en el curso Uso del Portafolio de Evaluaciones de Matemática en el Aula (Portafolio EMA), para Profesores de Matemática de Segundo Ciclo Básico, cuyo objetivo fue familiarizar a los docentes con una herramienta para el registro de evidencia de sus propias PEAA, la cual puedan utilizar para la reflexión pedagógica.

Correctores. Dos correctores fueron seleccionados a partir de una lista de correctores recomendados por los equipos de Docentemás (programa de evaluación del desempeño docente obligatorio para docentes empleados por el Estado de Chile) y del programa de Asignación a la Excelencia Pedagógica (AEP; programa chileno de certificación de excelencia docente voluntario para docentes del sistema municipal y particular subvencionado). Ambos ya habían sido correctores de los procesos de Docentemás y AEP y son profesores de matemática trabajando actualmente en aulas de segundo ciclo, con 13 y 20 años de experiencia, respectivamente.

Instrumentos

Portafolio EMA. Fue desarrollado en base al *QAS Notebook* (Martínez et al., 2012). Se realizó una traducción inicial al español y posteriormente una adaptación al subsector de matemática y al contexto chileno. El Portafolio EMA está constituido por tres carpetas. En la primera carpeta los profesores deben recolectar materiales de planificación de evaluaciones de la unidad en la cual trabajarán con el portafolio. En esta carpeta también deben contestar preguntas para contextualizar su curso. La segunda carpeta está constituida por 10 secciones correspondientes a 10 horas pedagógicas. En ellas cada profesor registra evaluaciones realizadas durante cada hora pedagógica, con material físico, oral o en el pizarrón. Junto con esto, los profesores adjuntan fotocopias de las evaluaciones realizadas por dos alumnos, uno de alto desempeño y uno de desempeño bajo el promedio. Por último, en la tercera carpeta los profesores adjuntan las evaluaciones a realizar después de las 10 horas pedagógicas y responden preguntas finales.

Rúbrica de corrección. Aborda nueve dimensiones relacionadas con las PEAA. Estas son: Establecimiento de objetivos claros de aprendizaje, Frecuencia de las evaluaciones, Variedad de las evaluaciones, Alineamiento de las evaluaciones con los objetivos de aprendizaje, Complejidad cognitiva, Razonamiento matemático, Involucramiento de los estudiantes en su propia evaluación, Retroalimentación del profesor al estudiante y Uso de la información para la toma de decisiones instruccionales. Las evidencias son puntuadas por los

correctores en una escala Likert de 5 puntos, donde 1 es *no presentado/realizado* y 5 *totalmente presentado/realizado.* Cada dimensión estuvo acompañada de una rúbrica que detalla el desempeño esperado para los niveles 1, 3 y 5, dejando los niveles 2 y 4 como posibilidad para que los correctores pudieran ubicar en esos niveles los desempeños intermedios. Para la evaluación de la dimensión 3, Variedad de las evaluaciones, no se utilizó la misma escala, sino que se optó por solicitar a los correctores llenar una lista de cotejo, identificando las evaluaciones realizadas cada día. Finalmente, se solicitó a los correctores que asignaran un puntaje global a cada portafolio.

Cuestionario a profesores. Constituido por un cuestionario de auto-reporte sobre PEAA en el cual se le solicita al docente información sobre la frecuencia de actividades de evaluación, el involucramiento de los estudiantes en su propia evaluación, el uso de la información de las evaluaciones para la toma de decisiones y la frecuencia de instancias de retroalimentación. Además, se incluyó una escala de auto-percepción sobre la propia capacidad o preparación para asesorar a otro profesor en aspectos relacionados con las PEAA.

Entrevistas finales con profesores y correctores. Para obtener información sobre posibles mejoras al instrumento y rúbrica y conocer la opinión de los docentes respecto del efecto del portafolio en su habilidad de evaluación y reflexión pedagógica, se realizaron entrevistas grupales o individuales con los profesores y con los correctores, solicitándoles también responder un cuestionario de reflexión.

Procedimiento

Cada profesor confeccionó dos portafolios durante el segundo semestre del año académico 2012. La muestra final de portafolios fueron 32 (15 profesores que realizaron dos portafolios y dos profesores que realizaron uno). Cada uno de los portafolios tomó a cada corrector alrededor de 60 minutos de corrección y fue corregido por ambos correctores (100% de doble corrección).

La aplicación del Portafolio EMA partió con una reunión grupal de capacitación de cuatro horas para los profesores sobre el uso del portafolio, la concepción de evaluaciones que se buscaba estudiar y los detalles del estudio. Al comenzar la reunión, los docentes contestaban el cuestionario a profesores. Después de la reunión, cada profesor elaboró los dos portafolios. Una de las investigadoras estuvo en constante contacto con los profesores en caso de que tuvieran dudas y se reunió con los que lo solicitaron para apoyar personalmente la terminación del primer portafolio. El proceso incluyó, luego de la primera aplicación, reuniones de evaluación grupal sobre la experiencia de construcción del portafolio. En estas reuniones se trabajó sobre la experiencia del uso del portafolio EMA, sin entregar retroalimentación respecto a las PEAA, ni del grupo ni de cada profesor. Las reuniones no fueron diseñadas como una instancia de enseñanza-aprendizaje sobre PEAA, sino solo respecto de la recolección de datos con el portafolio. Finalmente, luego de la segunda elaboración del portafolio, se realizaron entrevistas grupales e individuales, en las cuales los profesores contestaron nuevamente el cuestionario al profesor y realizaron una retroalimentación a la investigadora sobre el uso del portafolio EMA.

La capacitación de los correctores fue diseñada a partir de modelos de capacitación encontrados en la literatura en torno al tema de la evaluación de portafolios (Martínez et al., 2012; van der Schaaf, Stokking & Verloop, 2005). Antes de la primera capacitación se entregó a los correctores el Portafolio EMA para que ellos pudieran familiarizarse con el instrumento. Luego recibieron una capacitación inicial de cuatro horas de introducción al Portafolio EMA y al proyecto de investigación, con el objetivo de familiarizarlos con la concepción de evaluaciones con la que se estaba trabajando. En una segunda capacitación de cinco horas se revisó con ellos y en detalle cada una de las dimensiones de la rúbrica de evaluación, aclarando dudas y aunando criterios. Finalmente, en una tercera reunión de cinco horas se trabajó en la corrección de dos portafolios evaluados previamente por el equipo de investigación y expertos en la disciplina (*codificación maestra*). El primero fue evaluado de manera conjunta y el segundo, individualmente, contrastando los resultados para calibrar los criterios de evaluación y clarificar dudas.

Cada corrector evaluó 29 portafolios en jornadas de cuatro horas cada día, con un máximo de cuatro portafolios por día. Luego de la corrección de los dos primeros portafolios, al notar falta de acuerdo en algunas dimensiones, se realizó una reunión para trabajar las dimensiones conflictivas, aclarar dudas y lograr consenso. En la mitad del tiempo de corrección se realizó una reunión de cuatro horas de re-calibración, realizando una comparación de los puntajes que cada corrector asignó a un tercer portafolio de la codificación maestra.

Análisis

Para evaluar la confiabilidad entre jueces se realizó un análisis de correlación (Tinsley & Weiss, 2000). Dadas las limitaciones de este método en cuanto a su capacidad de captar diferencias en los promedios y dar luces sobre el impacto de distintas fuentes de error de medición (Tinsley & Weiss, 2000), se implementaron dos estudios de generalizabilidad. Los estudios de generalizabilidad entregan información más detallada sobre las fuentes de error de medición y obtienen coeficientes de confiabilidad de la corrección para la toma de decisiones relativas (ρ) y absolutas (Φ) (Brennan, 2001; Shavelson & Webb, 1991). Se utilizaron dos diseños distintos para investigar la influencia de dos posibles fuentes de error (llamadas *facetas*) identificadas por la literatura: corrector y ocasión. Los correctores pueden representar una fuente de error sistemática al no desempeñarse de manera óptima en la corrección de los portafolios, por ejemplo, al malinterpretar la rúbrica de puntuación. El momento en el cual se realiza la medición (llamado *ocasión*) también puede constituir una fuente de error sistemático, atribuible, por ejemplo, a distintos contenidos curriculares que se reflejan en los portafolios de un mismo profesor y aula.

Primero se realizó un diseño de *una faceta cruzada*, Profesor x Corrector (p*c), tomando en cuenta que cada corrector evaluó el portafolio de cada profesor participante. Sin embargo, en este primer diseño el análisis de generalizabilidad se realizó de manera separada tanto para el primer portafolio (n = 17) como para el segundo (n = 13). Se descompuso la varianza de cada una de las dimensiones en tres posibles fuentes de varianza: dos efectos principales, profesor (σ^2_p) y corrector (σ^2_c), y un componente de varianza residual que da cuenta de la interacción entre el profesor y el corrector y un error residual atribuible a diversas fuentes de varianzas no incluidas en el modelo ($\sigma^2_{pc,e}$). En el segundo diseño se utilizaron *dos facetas cruzadas*, Profesor x Corrector x Ocasión (p*c*o), tomando en cuenta que cada corrector evaluó los dos portafolios de cada profesor participante en los dos momentos de tiempo (ocasiones). Entonces, en este diseño se pudieron distinguir y comparar en un solo análisis de generalizabilidad tres efectos principales: la varianza verdadera entre los profesores (σ^2_p), un error de varianza proveída por los correctores (σ^2_c) y un error de varianza atribuido a la ocasión (σ^2_o). También se identificaron tres efectos de interacción: entre el profesor y el corrector (σ^2_{pc}), entre el profesor y la ocasión (σ^2_{po}) y entre el corrector y la ocasión (σ^2_{co}). Por último, había un componente de varianza atribuido a una interacción entre el profesor, corrector y ocasión, más fuentes de varianza no incluidos en el modelo ($\sigma^2_{pco,e}$). Para realizar estos análisis se utilizó el comando VARCOMP en SPSS 19, con el método estimador mínimo no cuadrático insesgado (EMNCI o MINQUE en inglés; Rao, 1971).

Con el objetivo de contrastar el auto-reporte de los profesores respecto de las PEAA con sus resultados en el Portafolio EMA, se llevó a cabo un análisis de correlación de las dimensiones del portafolio 1 con los indicadores del cuestionario de auto-reporte. Se realizó este contraste buscando observar la correlación del auto-reporte inicial con la evidencia de las PEAA del primer portafolio, evitando, así, posibles efectos de aprendizaje por el uso del portafolio EMA.

Para investigar la pregunta si los profesores participantes en el pilotaje experimentaran una posible mejora de su auto-percepción respecto de sus habilidades para realizar PEAA, se realizó una comparación de medias pre-post la elaboración de los dos portafolios, en base de las respuestas de los profesores a la escala de auto-percepción sobre la propia capacidad o preparación para asesorar a otro profesor en aspectos relacionados con las evaluaciones de aprendizaje. También se realizó un análisis cualitativo de contenido de las grabaciones de dos entrevistas grupales, del cual se detectaron algunas temáticas emergentes en cuanto al uso del Portafolio EMA y a su influencia en la reflexión sobre las PEAA.

Resultados

Análisis Descriptivo

Al analizar los datos descriptivos de la evaluación de los portafolios, en una escala donde 1 = *muy bajo* y 5 = *muy alto*, se observan promedios muy altos para la dimensión 2 y promedios bajos en las dimensiones 7, 8 y 9 (ver Tabla 1). Esto significa que los profesores de la muestra utilizan evaluaciones tanto formativas como sumativas con alta frecuencia, pero no hacen un uso adecuado de sus resultados para informar sus prácticas y entregar retroalimentaciones de calidad a sus estudiantes. Además, no se reportan suficientes instancias que permitan el involucramiento de los estudiantes en su propia evaluación.

Tabla 1
Estadísticos Descriptivos por Dimensión y Portafolio

Dimensión	Portafolio 1 (n = 17)		Portafolio 2 (n = 13)	
	Promedio	Desv. típica	Promedio	Desv. típica
1. Objetivos de aprendizaje	2,50	1,60	2,77	1,70
2. Frecuencia de las evaluaciones	4,68	0,54	4,46	0,86
3. Variedad de las evaluaciones	3,49	0,69	3,34	0,72
4. Alineamiento de las evaluaciones con los objetivos de aprendizaje	3,00	1,67	3,46	1,73
5. Complejidad cognitiva	3,79	1,04	3,92	0,98
6. Razonamiento matemático	2,62	1,33	3,04	1,22
7. Involucramiento de los estudiantes	1,76	1,02	1,38	0,64
8. Retroalimentación	1,82	1,00	1,65	1,06
9. Uso de la información	2,32	1,36	2,00	1,30
10. Prácticas de evaluación general	3,06	0,81	3,12	0,86

Correlación entre Correctores

Al observar la correlación entre los correctores (ver Tabla 2), se encuentra una alta correlación en las dimensiones 1 y 4, una correlación media en las dimensiones 8, 9 y 10 y una correlación baja en las dimensiones 2, 3, 5, 6 y 7.

Tabla 2
Correlación Entre Jueces por Dimensión y Portafolio

Dimensión	Portafolio 1	Portafolio 2	Promedio
1. Objetivos de aprendizaje	0,85	0,98	0,92
2. Frecuencia de las evaluaciones	0,14	0,25	0,20
3. Variedad de las evaluaciones	0,22	0,31	0,27
4. Alineamiento de las evaluaciones con los objetivos de aprendizaje	0,93	0,85	0,89
5. Complejidad cognitiva	0,03	0,59	0,31
6. Razonamiento matemático	0,30	0,21	0,26
7. Involucramiento de los estudiantes	0,41	0,06	0,24
8. Retroalimentación	0,70	0,54	0,62
9. Uso de la información	0,53	0,62	0,58
10. Prácticas de evaluación general	0,42	0,47	0,45
Promedio	0,45	0,49	0,47

Estudio de Generalizabilidad

En el primer estudio de generalizabilidad, con una faceta cruzada (Profesor x Corrector), se observa que en las dimensiones 1, 4 y 8 del primer portafolio un alto porcentaje de la varianza (> 53%) está explicado por la varianza verdadera entre los portafolios (ver Tabla 3). Estas dimensiones mantienen un alto porcentaje en el segundo portafolio, exceptuando la dimensión 8. Las dimensiones 2, 3, 5, 6, 7 y 10 en los portafolios 1 y 2 dan cuenta de un gran porcentaje de varianza atribuido principalmente a un error de medición inespecífico o a variables no incluidas en el modelo (> 55%). En las dimensiones 3, 6 y 9 en el primer portafolio se observa un porcentaje de varianza considerable atribuido al corrector.

Tabla 3
Estudio de Generalizabilidad de los Portafolios EMA por Dimensión (Diseño p x c)

Dimensión	Primer Portafolio EMA (n = 17)				Segundo Portafolio EMA (n = 13)			
	σ^2_{total} (100%)	σ^2_p	σ^2_c	$\sigma^2_{pc,e}$	σ^2_{total} (100%)	σ^2_p	σ^2_c	$\sigma^2_{pc,e}$
1. Objetivos de aprendizaje	2,63	73,3	_	26,7	3,04	95,0	1,3	3,8
2. Frecuencia de las evaluaciones	0,29	7,7	_	93,0	0,76	14,3	_	85,7
3. Variedad de las evaluaciones	0,53	9,6	22,3	68,1	0,60	14,1	29,1	56,8
4. Alineamiento de las evaluaciones con los objetivos de aprendizaje	2,88	85,7	1,4	12,9	3,10	73,1	_	26,9
5. Complejidad cognitiva	1,11	1,6	_	98,4	0,99	42,2	_	57,8
6. Razonamiento matemático	2,10	11,9	33,0	55,1	1,86	6,9	42,1	51,0
7. Involucramiento de los estudiantes	1,05	25,5	_	74,5	0,42	3,1	_	96,9
8. Retroalimentación	1,02	53,6	_	46,4	1,14	35,9	1,7	62,4
9. Uso de la información	2,22	24,5	31,8	43,7	1,95	33,1	26,6	40,3
10. Prácticas de evaluación general	0,68	26,9	_	73,1	0,78	30,6	_	69,4

Nota. Los espacios vacíos corresponden a componentes de varianza igual a cero o negativos.
σ^2_{total} = Varianza total; σ^2_p = Varianza atribuible al profesor; σ^2_c = Varianza atribuible al corrector; $\sigma^2_{pc,e}$ = Varianza error (interacción entre el profesor y el corrector, más el error inespecífico).

En la Tabla 4, en el diseño con una faceta se puede observar que los coeficientes de generalizabilidad para la toma de decisiones absolutas (Φ) son bajos para casi todas las dimensiones, exceptuando las dimensiones 1 y 4, que presentan un coeficiente medio-alto, y la dimensión 8, que presenta un coeficiente moderado. Estos resultados dan cuenta de que para las dimensiones 1, 4 y 8 un gran porcentaje de la varianza corresponde a varianza verdadera del profesor. Al comparar los coeficientes ρ y Φ, se puede observar poca varianza entre ellos, ya que existe poca o nula varianza atribuida al corrector.

En el estudio con dos facetas (Tabla 4) se encuentran coeficientes medio-altos para la dimensión 4, al igual que en el estudio de una faceta. En comparación con el estudio de una faceta, se observa que el coeficiente Φ disminuye para las dimensiones 1 y 8, pasando de un coeficiente moderado a pequeño, lo que daría cuenta de una disminución de la confiabilidad de estas dimensiones para la toma de decisiones de carácter absoluto. En contraste, las dimensiones 5, 6 y 7 aumentan los coeficientes Φ, aunque estos siguen siendo pequeños. Las dimensiones 2, 3 y 10 obtuvieron un coeficiente igual a cero, ya que toda la varianza es atribuida a diversas fuentes de error, presentando un 0% de varianza verdadera.

En el diseño del estudio de generalizabilidad con dos facetas cruzadas (Profesor x Corrector x Ocasión), los resultados de este estudio (ver Tabla 5) dieron cuenta de un porcentaje de varianza pequeño de los correctores, salvo en las dimensiones 3, 6 y 9, en las que variaron entre 20 y 30%. Para el componente de varianza atribuible a la ocasión se observó un efecto de varianza casi nula, lo que da cuenta de que el momento en el tiempo en este estudio no tiene una influencia directa y uniforme en todos los profesores. Se observaron porcentajes de varianza pequeños a moderados en la interacción entre el profesor y la ocasión, exceptuando la dimensión 1, en que se encontró un alto porcentaje. Esto muestra que en la dimensión 1 la ocasión tiene una influencia en el desempeño de los profesores al hacer el portafolio, pero esta no es igual para todos los profesores, dependiendo del momento en el año en que fueron recolectados o la unidad pedagógica abordada con el portafolio. Se observó una varianza verdadera pequeña para todas las dimensiones, exceptuando la dimensión 4. Por último, se encontraron moderados a grandes porcentajes de varianza residual en casi todas las dimensiones.

Tabla 4
Coeficientes de Generalizabilidad

Dimensión	1 faceta[a]		2 facetas	
	ρ	φ	ρ	φ
1. Objetivos de aprendizaje	0,85	0,85	0,52	0,34
2. Frecuencia de las evaluaciones	0,14	0,14	0,00	0,00
3. Variedad de las evaluaciones	0,22	0,17	0,00	0,00
4. Alineamiento de las evaluaciones con los objetivos de aprendizaje	0,93	0,92	0,87	0,87
5. Complejidad cognitiva	0,03	0,03	0,32	0,30
6. Razonamiento matemático	0,30	0,21	0,55	0,39
7. Involucramiento de los estudiantes	0,41	0,41	0,44	0,44
8. Retroalimentación	0,70	0,70	0,41	0,39
9. Uso de la información	0,53	0,39	0,35	0,23
10. Prácticas de evaluación general	0,42	0,42	0,00	0,00
Promedio	0,45	0,42	0,35	0,30

[a] Se utilizaron los datos del Portafolio 1.

Tabla 5
Estudio de Generalizabilidad de los Portafolios EMA por Dimensión (Diseño p x c x o)

Dimensión	Componente de la varianza (%)							
	σ^2_{total} (100%)	σ^2_p	σ^2_c	σ^2_o	σ^2_{pc}	σ^2_{po}	σ^2_{co}	$\sigma^2_{pco,e}$
1. Objetivos de aprendizaje	2,78	19,3	_	_	_	65,5	4,5	10,7
2. Frecuencia de las evaluaciones	0,64	_	0,5	2,5	17,4	21,4	_	58,2
3. Variedad de las evaluaciones	0,84	_	21,6	_	23,9	15,5	_	38,9
4. Alineamiento de las evaluaciones con los objetivos de aprendizaje	3,00	68,5	0,1	_	_	10,0	_	21,3
5. Complejidad cognitiva	1,51	10,6	3,0	_	_	5,1	_	81,3
6. Razonamiento matemático	3,54	17,2	21,7	4,4	_	_	_	56,7
7. Involucramiento de los estudiantes	0,99	16,2	_	_	_	_	1,0	82,8
8. Retroalimentación	1,07	17,3	_	_	6,0	17,6	7,2	51,9
9. Uso de la información	2,17	10,8	29,8	1,0	9,2	13,1	_	36,1

Nota. Los espacios vacíos corresponden a componentes de varianza igual a cero o negativos.
σ^2_{total} = Varianza total; σ^2_p = Varianza atribuible al profesor; σ^2_c = Varianza atribuible al corrector; σ^2_o = Varianza atribuible a la ocasión; σ^2_{pc} = Varianza atribuible a la interacción entre profesor y corrector; σ^2_{po} = Varianza atribuible a la interacción entre profesor y ocasión; σ^2_{co} = Varianza atribuible a la interacción entre corrector y ocasión; $\sigma^2_{pco,e}$ = Varianza error (interacción entre el profesor, el corrector y la ocasión, más el error inespecífico).

Correlación entre Desempeño y Auto-Reporte

Al comparar los puntajes del Portafolio 1 con el auto-reporte de los profesores sobre sus PEAA (ver Tabla 6), no se observan correlaciones significativas directas, con la excepción de la dimensión 3. Estos resultados, de acuerdo a lo esperado, muestran que la percepción de los docentes respecto de sus prácticas no necesariamente es consistente con la evidencia física registrada en los portafolios, evaluados por los correctores según la rúbrica de puntuación. Incluso la dimensión 4 muestra una correlación inversa entre el auto-reporte docente y la evidencia de los portafolios.

Tabla 6
Correlación Entre Auto-Reporte Docente y Resultados en el Portafolio Según Dimensión

Dimensión	Correlación
1. Objetivos de aprendizaje	0,281
2. Frecuencia de las evaluaciones	0,022
3. Variedad de las evaluaciones	0,668**
4. Alineamiento de las evaluaciones con los objetivos de aprendizaje	-0,429*
5. Complejidad cognitiva	-0,108
6. Razonamiento matemático	0,166
7. Involucramiento de los estudiantes	-0,240
8. Retroalimentación	0,185
9. Uso de la información	-0,242

**$\rho < 0,01$; *$\rho < 0,05$

Auto-Percepción de Habilidades Evaluativas

En el cuestionario docente la escala de auto-percepción de habilidades evaluativas alcanzó un coeficiente alfa de Cronbach de 0,93, lo cual corresponde a una consistencia interna alta de esta escala. Al comparar las medias antes y después de la realización de los portafolios, no se encontraron diferencias en la dirección esperada, dando cuenta de un aumento significativo solo en la dimensión referente a cuán capacitados se perciben los participantes para asesorar a otro profesor en desarrollar planificaciones de evaluación para un curso o unidad, $F(1, 27) = 4,36$, $p = 0,046$, 95% IC [-1,048, -0,009], $\eta^2 = 0,139$.

También se realizó un análisis cualitativo de dos entrevistas grupales, en las cuales se detectaron algunas temáticas emergentes que se orientan al trabajo con el portafolio como un instrumento de recolección de datos y a la influencia del uso del portafolio en la reflexión docente. Respecto de la estructura del portafolio, se comentó que involucraba un número de clases adecuado, que no lo hacía abrumador, pero que al mismo tiempo obligaba a los docentes a plasmar la realidad de sus prácticas, ya que, como decía un profesor, "*en 10 días tampoco puedes verbalizar puras mentiras. Y creo que una fortaleza del portafolio es eso, el tema de las evidencias* (...) *nos fuerza un poco a la verdad, a no mentir*". Surgió el tema de la dificultad de reportar las clases por hora pedagógica, ya que eso los hacía dividir artificialmente una clase que en sí misma es una unidad con inicio, desarrollo y cierre. También se comentó lo difícil que les resultó a algunos el uso de las calcomanías de colores para identificar la evidencia de los alumnos con alto y bajo rendimiento.

Respecto de la influencia del portafolio en las prácticas pedagógicas, surgió fuertemente el rol que tuvo en mostrar a los profesores nuevas instancias de evaluación de aprendizajes que antes no tomaban en consideración y que ahora, al considerarlas como tales, quieren hacerlas mejor. Una profesora reportaba su pensamiento mientras hacía el portafolio diciendo: "*¡Ah! ¡¿Esto es evaluación?! Entonces lo voy a hacer mejor, porque voy a estar evaluando. No lo voy a hacer así no más*". Los profesores también reportaron que al utilizar el portafolio se dieron cuenta de que no contaban con evidencia física que respaldara el aprendizaje de los alumnos, que no prestaban tanta atención al trabajo en clase de los alumnos con bajo rendimiento como sí al de los de alto rendimiento y que utilizaban sistemáticamente las mismas herramientas de evaluación de manera indiferente a los contenidos trabajados. Por último, los profesores reportaron un enriquecimiento de sus reflexiones pedagógicas debido al uso del Portafolio EMA. Una profesora indicó: "*Yo creo que una de las fortalezas es que te da el espacio para la reflexión docente y eso te genera oportunidades de mejora para tomar decisiones*".

Dando respaldo a la evidencia reportada en las entrevistas grupales, en el cuestionario final de los profesores el 73% comentó que esta instancia fue un gran aporte para su desarrollo profesional, destacando como importante su cercanía con la práctica y el aporte para reflexionar sobre su desempeño docente.

Discusión

Limitaciones del Estudio

El hecho de que se utilizó una muestra pequeña y no probabilística impide generalizar los resultados descriptivos a otros profesores chilenos de matemática de segundo ciclo básico. Se trabajó con profesores altamente motivados y con especial interés en la temática bajo estudio, lo cual claramente tiene una influencia en la representatividad de los resultados. Sin embargo, se trató de un estudio de adaptación y pilotaje y se cuidó la participación de profesores de colegios que representan distintos contextos socioeconómicos.

Descripción de las PEAA de Profesores de Matemática de Segundo Ciclo de Enseñanza Básica en Chile

En los profesores participantes del presente estudio existe un uso frecuente de actividades de evaluación. Sin embargo, existe un bajo involucramiento de los estudiantes en su propia evaluación, al igual que una escasa retroalimentación a los alumnos por parte de los profesores. Este último hallazgo es de gran relevancia, puesto que en la literatura se ha visto que estos dos aspectos de las PEAA son especialmente relevantes para el desarrollo del aprendizaje auto-regulado en los estudiantes (Brookhart, 2013). En numerosas investigaciones se ha profundizado en el impacto que tiene la retroalimentación en los procesos de aprendizaje de los alumnos (Hattie & Timperley, 2007; Wiliam et al., 2004), la cual puede modificar la motivación de los estudiantes frente al aprendizaje (Stefanou & Parkes, 2003) y la forma en la cual los alumnos enfrentan las PEAA, como instancias de evaluación de habilidades o como instancias de desarrollo y aprendizaje (Stiggins, 2001).

También se observó un bajo uso de la información obtenida a través de las evaluaciones realizadas, lo cual es consistente con los resultados de la Evaluación Docente en Chile, que manifiesta bajos resultados de los profesores en el indicador *Reflexión a partir de los resultados de la evaluación*, incluso en los docentes que muestran un desempeño destacado (MINEDUC, 2012), lo que, a su vez, podría repercutir en las decisiones instruccionales.

Validación Preliminar del Portafolio EMA Como Herramienta Confiable Para Medir las Prácticas Evaluativas

Realizar evaluaciones de las prácticas docentes de manera válida y confiable es tarea ardua, debido a la complejidad de los conceptos a medir y la gran cantidad de variables influyentes (Romo et al., 2014). Incluso en estudios con muchos recursos disponibles, como el Measures of Effective Teaching (MET; Kane & Staiger, 2012), el control de las fuentes de error de medición es difícil de lograr. A través de los estudios de generalizabilidad del portafolio de la Evaluación Docente en Chile, se han encontrado porcentajes de varianza verdadera entre los profesores que varían entre un 25% y un 50%, mientras que los porcentajes atribuibles a los correctores son más bien pequeños (entre 3 y 10%). De manera similar, el estudio de validación del QAS Notebook muestra porcentajes de varianza verdadera entre 27,6% y 52,5%, dependiendo de la dimensión evaluada (Martínez et al., 2012). Los coeficientes de confiabilidad encontrados para el módulo escrito del portafolio de la Evaluación Docente en Chile son 0,73 para la toma de decisiones relativas (ρ) y 0,43 para la toma de decisiones absolutas (Φ; Taut, Santelices & Manzi, 2011).

Si bien algunas dimensiones del Portafolio EMA alcanzan porcentajes de varianza verdadera y coeficientes de generalizabilidad similares a los resultados de esos estudios, de manera global el instrumento y la rúbrica de corrección del Portafolio EMA muestran necesidad de ser perfeccionados. Los resultados encontrados en este estudio se podrían deber a poca claridad y adecuación del Portafolio EMA en algunas de sus dimensiones, la escasa varianza entre los portafolios en algunas dimensiones, la rúbrica para esas dimensiones y la capacitación de los profesores y/o de los correctores. La familiaridad de los correctores con la rúbrica de evaluación, la cantidad de tiempo dedicado a la capacitación y el hecho de que el portafolio y la rúbrica de evaluación se encuentran en un estado de pilotaje son factores importantes que aumentan los porcentajes de la varianza atribuibles a error (Randel & Clark, 2013).

Por ejemplo, es interesante analizar por qué las dimensiones Objetivos de aprendizaje y Alineación de las evaluaciones con los objetivos de aprendizaje fueron las dimensiones con la mejor generalizabilidad y en las que los correctores constituyeron una fuente de error casi nula. Estos resultados se podrían atribuir a la familiaridad de los correctores con estas dimensiones, ya que existen de manera similar en los portafolios de la

Evaluación Docente (Flotts & Arbazúa, 2011) y de la Asignación de Excelencia Pedagógica en Chile (Argüelles, Saragoni & Castillo, 2015). Las dimensiones sobre conceptos más novedosas, como Involucramiento de los estudiantes y Complejidad cognitiva, con menos familiaridad de los correctores, fueron también las menos generalizables.

Por otro lado, comparando la capacitación de los correctores de este estudio y su conocimiento de la rúbrica de evaluación con los del programa PACT (Pecheone & Chung, 2006), la Evaluación Docente en Chile (Flotts & Abarzúa, 2011), el programa del NBPTS (Jaeger, 1998) y el estudio MET (Kane & Staiger, 2012), se observa que estos realizan arduos procesos de selección y capacitación de los correctores, debiendo estos realizar pruebas de selección y certificación, demostrando competencias mínimas para el trabajo y un buen manejo de las rúbricas. Surge, entonces, la importancia de un proceso exhaustivo de selección y capacitación, analizando en profundidad las representaciones mentales de los correctores, sus procesos cognitivos y su adecuada comprensión de las dimensiones evaluativas de la rúbrica (García, Torres & Leyton, 2013). En el presente contexto parece especialmente importante incluir en el proceso de selección evidencia de la experticia de los correctores respecto de las PEAA, ya que se ha visto que esta es fundamental para el proceso de evaluación y asignación de un puntaje a través del uso de rúbricas de evaluación (Randel & Clark, 2013).

El portafolio como herramienta de evaluación ofrece una mirada más completa y global del desempeño docente e incluye una mayor cantidad de fuentes de información, lo que hace que la labor interpretativa de los correctores sea una parte fundamental (van der Schaaf et al., 2005). Los correctores, antes de evaluar, se hacen una imagen mental de la evidencia, para luego —influidos por su comprensión de los criterios de evaluación, las experiencias previas, el contexto, las condiciones del proceso de corrección y la importancia que le asignan a diferentes evidencias del portafolio— asignan un puntaje para cada dimensión (Heller, Sheingold & Myford, 1998). Entre todos los aspectos a trabajar, surge, por lo tanto, la necesidad de orientar la capacitación hacia las representaciones cognitivas que los correctores tienen de las dimensiones evaluadas (García et al., 2013). En su investigación realizada con seis correctores y doble corrección de 18 portafolios, van der Schaaf et al. (2005) encontraron una alta correlación entre las representaciones cognitivas de los correctores y la confiabilidad de los puntajes de corrección. Su estudio también hace notar la dificultad de modificar estas representaciones, ya que ellas han sido adquiridas a través de años de experiencia docente y en las experiencias previas como correctores.

Hallazgos Sobre la Autopercepción de los Docentes Respecto de las PEAA

La baja correlación encontrada entre los resultados del cuestionario de auto-reporte sobre las PEAA de los profesores y las dimensiones evaluadas en el portafolio indica que la información de ambos instrumentos se superpone de manera limitada, resaltando la importancia de evaluar estas capacidades en base de evidencia directa.

Al comparar los resultados de la escala de autopercepción de la propia capacidad para realizar PEAA antes y después del proyecto, se encontró un aumento general de los puntajes de los profesores participantes. Sin embargo, la diferencia fue significativa únicamente para la dimensión que evaluaba la autopercepción docente respecto de su capacidad para desarrollar planificaciones de evaluación para un curso o unidad. Esto puede deberse al pequeño tamaño de la muestra y a los altos puntajes iniciales. De hecho, en el presente estudio se pudo haber generado un sesgo en la muestra de profesores, ya que, al ser una actividad voluntaria de alta exigencia, es probable que los docentes con una percepción negativa de sus habilidades para realizar PEAA se hayan abstenido de participar (Tschannen-Moran, Hoy & Hoy, 1998). De esta manera, queda pendiente la evaluación de un posible impacto del Portafolio EMA en la autopercepción de los docentes que tienen baja percepción de sus habilidades para realizar PEAA.

El Potencial del Portafolio EMA Como Herramienta de Reflexión Docente

A través de la evaluación cualitativa, tanto los profesores como los correctores indicaron que la participación en el pilotaje fue un gran aporte a sus prácticas docentes, valorando lo cercano que fue de su práctica cotidiana y destacando la instancia como contribución a su reflexión pedagógica. Esta evaluación de los profesores es consistente con la evidencia encontrada por la Evaluación Docente en Chile (Taut & Sun, 2014), el NBPTS (Sato, Wei & Darling-Hammond, 2008; Wolf & Taylor, 2008) y el programa PACT, en los cuales los profesores indican que la realización de los portafolios tuvo un impacto en el desarrollo de sus habilidades, su capacidad para reflexionar y sus PEAA de sus alumnos (Pecheone & Chung, 2006). En la literatura se habla del uso

del proceso de evaluación por sobre el uso de los resultados que este genera (Patton, 2008). Este fenómeno es especialmente relevante cuando los evaluados están involucrados en el proceso de evaluación, como es el caso de la implementación de un portafolio como herramienta de evaluación. Así, la evaluación no solo genera resultados que pueden constituir un impulso para el evaluado y otros implicados, sino también significa en sí misma una instancia de reflexión.

Conclusiones

La presente investigación es un aporte al escaso estudio existente sobre las PEAA en Chile. Se trata del primer estudio piloto de un portafolio específicamente enfocado en las PEAA, las cuales corresponden a un aspecto relevante para distinguir la calidad de las prácticas docentes y que han recibido poca atención en el estudio de la calidad de las prácticas docentes.

Respecto de la pregunta formulada sobre la confiabilidad del Portafolio EMA en cuanto a su función de evaluación docente, los análisis realizados dan cuenta de una baja generalizabilidad de los resultados del Portafolio EMA para la toma de decisiones relativas y absolutas sobre el desempeño docente en esta área. Esto refleja que en el estado actual el Portafolio EMA no sería una herramienta confiable para medir las PEAA de docentes de matemática. Sin embargo, los cuestionarios de auto-reporte docente sobre las PEAA no parecen constituir una alternativa válida. Según lo esperado, y en respuesta a la segunda pregunta guía del estudio, se encuentra una baja correlación entre el auto-reporte de los docentes respecto de sus PEAA y la evidencia empírica directa recolectada por el Portafolio EMA. La literatura internacional confirma la complejidad de medir prácticas evaluativas en el aula de manera válida y confiable. De esta manera se indica, por un lado, la necesidad de perfeccionar el instrumento y la rúbrica, manteniendo algunas dimensiones y modificando otras y, por otro lado, desarrollar un más exhaustivo y efectivo proceso de selección y capacitación de los correctores. Solo a partir de estas modificaciones y un segundo estudio de pilotaje y validación se podría pretender utilizar el Portafolio EMA como herramienta para la evaluación de las PEAA en Chile.

Por último, el estudio encuentra indicios de carácter cualitativo y preliminar de que el Portafolio EMA constituye una instancia de reflexión pedagógica útil para los docentes que lo aplicaron en su práctica en aula enseñando matemática en el segundo ciclo de la enseñanza básica. Si bien no se pudo constatar un aumento en la autopercepción docente, con la excepción de la dimensión de la habilidad para desarrollar planificaciones de evaluación para un curso o unidad, se detecta un posible sesgo en la muestra que hace necesario, en futuras investigaciones, indagar más en el potencial promisorio de esta herramienta para el desarrollo profesional docente.

Referencias

Argüelles, M., Saragoni, C. & Castillo, M. Á. (2015). Sistema de evaluación e instrumentos de medición del programa AEP. En B. Rodríguez, J. Manzi, C. Peirano, R. González & D. Bravo (Eds.), *Reconociendo el mérito docente. Programa de asignación de excelencia pedagógica 2002-2014* (pp. 24-28). Santiago, Chile: Pontificia Universidad Católica de Chile, Escuela de Psicología, MIDE UC. Extraído de http://www.mideuc.cl/wp-content/uploads/2015/01/Libro-digital-Reconociendo-el-m%C3%A9rito-docente.-Programa-de-Asignaci%C3%B3n-de-Excelencia-Pedag%C3%B3gica-2002-2014.pdf

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5,* 7-74. doi:10.1080/0969595980050102 Extraído de http://search.proquest.com.ezproxy.puc.cl/docview/204052267

Bond, L., Smith, T., Baker, W. & Hattie, J. (2000). *The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study*. Washington, DC: National Board for Professional Teaching Standards.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Brookhart, S. M. (2013). Grading. En J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 35-54). Thousand Oaks, CA: SAGE.

Campbell, C. (2013). Research on teacher competency in classroom assessment. En J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 71-84). Thousand Oaks, CA: SAGE.

Chile, Ministerio de Educación (2012). *Resultados evaluación docente 2011*. Santiago, Chile: Autor, Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas. Extraído de http://www.docentemas.cl/docs/Resultados_Evaluacion_Docente_2011_270312.pdf

Flotts, M. P. & Abarzúa, A. (2011). El modelo de evaluación y los instrumentos. En J. Manzi, R. González & Y. Sun (Eds.), *La evaluación docente en Chile* (pp. 35-61). Santiago, Chile: Pontificia Universidad Católica de Chile, Escuela de Psicología, MIDE UC.

García, M. R., Torres, P. & Leyton, C. (2013). Representaciones cognitivas involucradas en la corrección de portafolios docentes. *Pensamiento Educativo: Revista de Investigación Educacional Latinoamericana, 50*(1), 21-39. doi:10.7764/PEL.50.1.2013.3 Extraído de http://pensamientoeducativo.uc.cl/index.php/pel/article/view/552

Gitomer, D. (2008). Reliability and NBPTS assessment. En L. Ingvarson & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards* (pp. 231-253). Oxford, Reino Unido: Elsevier.

Goldhaber, D. & Anthony, E. (2004). *Can teacher quality be effectively assessed?* Washington, DC: Urban Institute. Extraído del sitio Web del Urban Institute: http://www.urban.org/url.cfm?ID=410958

Hattie, J. (2008). Validating the specification of standards for teaching: Applications to the National Board for Professional Teaching Standards' assessments. En L. Ingvarson & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards* (pp. 93-111). Oxford, Reino Unido: Elsevier.

Hattie, J. & Clinton, J. (2008). Identifying accomplished teachers: A validation study. En L. Ingvarson & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards* (pp. 313-344). Oxford, Reino Unido: Elsevier.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77,* 81-112. doi:10.3102/003465430298487

Heller, J. I., Sheingold, K. & Myford, C. M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment, 5,* 5-40. doi:10.1207/s15326977ea0501_1

Jaeger, R. M. (1998). Evaluating the psychometric qualities of the National Board for Professional Teaching Standards' assessments: A methodological accounting. *Journal of Personnel Evaluation in Education, 12,* 189-210. doi:10.1023/A:1008085128230

Kane, T. J. & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation. Extraído de http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Kennedy, M. (Ed.) (2010). *Teacher assessment and the quest for teacher quality: A handbook*. San Francisco, CA: John Wiley & Sons.

Manzi, J., González, R. & Sun, Y. (Eds.) (2011). *La evaluación docente en Chile*. Santiago, Chile: Pontificia Universidad Católica de Chile, Escuela de Psicología, MIDE UC.

Martínez, J. F. (2013). Combinación de mediciones de la práctica y el desempeño docente: consideraciones técnicas y conceptuales para la evaluación docente. *Pensamiento Educativo: Revista de Investigación Educacional Latinoamericana, 50*(1), 4-20. doi:10.7764/PEL.50.1.2013.2

Martínez, J. F., Borko, H., Stecher, B., Luskin, R. & Kloser, M. (2012). Measuring classroom assessment practice using instructional artifacts: A validation study of the QAS Notebook. *Educational Assessment, 17,* 107-131. doi:10.1080/10627197.2012.715513

McMillan, J. H. (2013). Why we need research on classroom assessment. En J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 3-16). Thousand Oaks, CA: SAGE.

Montecinos, C., Rittershaussen, S., Solís, M. C., Contreras, I. & Contreras, C. (2010). Standards-based performance assessment for the evaluation of student teachers: A consequential validity study. *Asia-Pacific Journal of Teacher Education, 38,* 285-300. doi:10.1080/1359866X.2010.515941

National Research Council (2008). *Assessing accomplished teaching: Advanced-level certification programs*. Committee on Evaluation of Teacher Certification by the National Board for Professional Teaching Standards. M. D. Hakel, J. A. Koening & S. W. Elliott (Eds.). Washington, DC: National Academies Press.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4ª ed.). Thousand Oaks, CA: SAGE.

Pecheone, R. L. & Chung, R. R. (2006). Evidence in teacher education: The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education, 57,* 22-36. doi:10.1177/0022487105284045

Pool, J. E., Ellett, C. D., Schiavone, S. & Carey-Lewis, C. (2001). How valid are the National Board of Professional Teaching Standards assessments for predicting the quality of actual classroom teaching and learning? Results of six mini case studies. *Journal of Personnel Evaluation in Education, 15,* 31-48. doi:10.1023/A:1011152101776

Randel, B., Beesley, A. D., Apthorp, H., Clark, T. F., Wang, X., Cicchinelli, L. F. & Williams, J. M. (2011). *Classroom assessment for student learning: The impact on elementary school mathematics in the Central Region* (NCEE 2011-4005). Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance/U.S. Department of Education.

Randel, B. & Clark, T. (2013). Measuring classroom assessment practices. En J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 145-163). Thousand Oaks, CA: SAGE.

Rao, C. R. (1971). Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis, 1,* 445-456. doi:10.1016/0047-259X(71)90019-4

Romo, F., Treviño, E. & Villalobos, C. (2014). *Capacidades de enseñanza en el sistema escolar: análisis de la evaluación docente en Chile* (Informes para la Política Educativa N° 3). Santiago, Chile: Universidad Diego Portales, Centro de Políticas Comparadas de Educación.

Santelices, M. V. & Taut, S. (2011). Convergent validity evidence regarding the validity of the Chilean standards-based teacher evaluation system. *Assessment in Education: Principles, Policy & Practice, 18,* 73-93. doi:10.1080/0969594X.2011.534948

Santiago, P., Benavides, F., Danielson, C., Goe, L. & Nusche, D. (2013). Teacher evaluation in Chile 2013 (OECD Reviews of Evaluation and Assessment in Education). Paris, Francia: OECD. doi:10.1787/9789264172616-en

Sato, M., Wei, R. C. & Darling-Hammond, L. (2008). Improving teachers' assessment practices through professional development: The case of National Board certification. *American Educational Research Journal, 45,* 669-700. doi:10.3102/0002831208316955 Extraído de http://www.jstor.org/stable/27667147

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: SAGE.

Stefanou, C. & Parkes, J. (2003). Effects of classroom assessment on student motivation in fifth-grade science. *The Journal of Educational Research, 96,* 152-162. doi:10.1080/00220670309598803

Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10*(1), 7-12. doi:10.1111/j.1745-3992.1991.tb00171.x

Stiggins, R. J. (2001). *Student-involved classroom assessment* (3ª ed.). Upper Saddle River, NJ: Prentice Hall.

Taut, S., Santelices, V. & Manzi, J. (2011). Estudios de validez de la evaluación docente. En J. Manzi, R. González & Y. Sun (Eds.), *La evaluación docente en Chile* (pp. 157-175). Santiago, Chile: Pontificia Universidad Católica de Chile, Escuela de Psicología, MIDE UC.

Taut, S. & Sun, Y. (2014). The development and implementation of a national, standards-based, multi-method teacher performance assessment system in Chile. *Education Policy Analysis Archives, 22*(71), 1-33. doi:10.14507/epaa.v22n71.2014 Extraído de http://epaa.asu.edu/ojs/article/view/1468/1313

Tinsley, H. E. A. & Weiss, D. J. (2000). Interrater reliability and agreement. En H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95-124). San Diego, CA: Academic Press.

Tschannen-Moran, M., Hoy, A. W. & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68,* 202-248. doi:10.3102/00346543068002202 Extraído de http://search.proquest.com.ezproxy.puc.cl/docview/214114604/abstract?accountid=16788

Tucker, P. D., Stronge, J. H., Gareis, C. R. & Beers, C. S. (2003). The efficacy of portfolios for teacher evaluation and professional development: Do they make a difference? *Educational Administration Quarterly, 39,* 572-602. doi:10.1177/0013161X03257304

van der Schaaf, M., Stokking, K. & Verloop, N. (2005). Cognitive representations in raters' assessment of teacher portfolios. *Studies in Educational Evaluation, 31,* 27-55. doi:10.10 l 6/j.stueduc.2005.02.005

Wiliam, D., Lee, C., Harrison, C. & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice, 11,* 49-65. doi:10.1080/0969594042000208994

Wolf, K. & Taylor, G. (2008). Effects of the National Board for Professional Teaching Standards: Certification process on teachers' perspectives and practices. En L. Ingvarson & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards* (pp. 381-412). Oxford, Reino Unido: Elsevier.

Fecha de recepción: Diciembre de 2013.
Fecha de aceptación: Junio de 2015.