PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2013, 50(1), 21-39

Pontificia Universidad Católica de Chile
Facultad de Educación

Representaciones cognitivas involucradas en la corrección de portafolios docentes

Cognitive Representations Involved in the Assessment of Teachers' Portfolios

María Rosa García, Pablo E. Torres y Carolina Leyton

Centro de Medición MIDE UC, Pontificia Universidad Católica de Chile

Resumen

En el Sistema de Evaluación del Desempeño Profesional Docente en Chile, uno de los instrumentos centrales para evaluar el desempeño docente es el portafolio, el cual es corregido por profesores de aula, seleccionados y capacitados para realizar esta tarea. Esta corrección es un proceso central del sistema, ya que incide en la validez de las inferencias posibles de extraer de los resultados obtenidos. Es por ello que comprender el razonamiento de los correctores al evaluar y corregir la evidencia se ha vuelto un eje de interés. En este estudio se presentan resultados asociados a las representaciones cognitivas que emplean los correctores al corregir evidencia del portafolio y las relaciones que existen entre la cantidad y la cualidad de estas con la precisión de la corrección. Seis profesores corrigieron seis entradas (evidencias) del portafolio escrito utilizando pautas de evaluación (rúbricas) y el método de *pensamiento en voz alta*. Los resultados muestran que la precisión evaluativa se ve afectada por la cualidad de las representaciones utilizadas más que por su cantidad. Asimismo, los resultados orientan sobre los tipos de representaciones y conceptos que resultan más efectivos al representarse la evidencia entregada por los evaluados y al representarse la pauta de corrección.

Palabras clave: evaluación docente, correctores, portafolio, representaciones mentales, rúbricas de evaluación

Correspondencia a:
María Rosa García, Centro de Medición MIDE UC, Pontificia Universidad Católica de Chile
Vicuña Mackenna 4860, Macul, Santiago, Chile
Correo electrónico: mrgarci1@uc.cl

El presente estudio se realizó en el Centro de Medición MIDE UC, el cual ha sido contratado por el Ministerio de Educación (MINEDUC) para prestar asesoría técnica para la implementación del Sistema de Evaluación del Desempeño Profesional Docente (desde 2003 a la fecha). La coordinación nacional y responsabilidad global del sistema de evaluación reside en el MINEDUC a través del Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas (CPEIP). Las opiniones emitidas por los investigadores son de su exclusiva responsabilidad y no representan ni comprometen la opinión del MINEDUC ni de CPEIP.

ISSN: 0719-0409 DDI: 203.262, Santiago, Chile
doi:10.7764/PEL.50.1.2013.3

Abstract

In the Chilean National Teacher Evaluation System, one of the central instruments used to assess teachers' performance is the Portfolio, which is reviewed and scored by school teachers trained and supervised to do so. The rating process is central to the system, because it impacts the extent to which valid inferences can be made from the assessment outcomes. Thus, understanding raters cognition has become an interesting area of research. This study shows results from cognitive representations used by raters during the Portfolio rating process. It specifically focuses on the relationship between the accuracy reached in the assessment and the quantity and quality of mental representations. Six teachers rated six portfolio entries using evaluation rubrics and following the think aloud procedure. The results show that it is not the amount but the quality of mental representations that relates to accuracy in the assessments of written portfolio. The results provide insights into the types of representations and concepts that seem to be more effective when the raters represent both, the evidence given by the evaluated teachers and the rubrics with the assessment criteria.

Keywords: teacher evaluation, raters, portfolio, mental representations, scoring rubrics

Las evaluaciones o mediciones que se basan en el juicio de personas centran gran parte de su validez en los procesos cognitivos de sus jueces, afectando las inferencias que se realizan a partir de estos (Crisp, 2012). La evaluación de portafolio no está exenta de este problema. Las actividades cognitivas de los correctores han demostrado ser una importante fuente de varianza en su medición (van der Shaaf, Stokking, & Verloop, 2005). Ello probablemente se deba a que los portafolios son instrumentos muy ligados a aspectos personales y de contexto, lo cual hace su evaluación una tarea altamente demandante y compleja, aún después de entrenamiento (van der Shaaff et al., 2005). No obstante, las actividades cognitivas desplegadas en la evaluación de portafolios han mostrado ser influenciables y aprendibles, por medio de entrenamiento, aumentando la precisión y disminuyendo el error de puntuación (Jonassen, 2000; Lievens, 2001; Woehr & Huffcutt, 1994). Es por ello que los estudios que intentan entender la actividad cognitiva de los correctores de portafolio son relevantes para el aseguramiento de la validez de la corrección, en la medida en que pueden incidir en sus sistemas de entrenamiento y monitoreo.

Dentro del contexto chileno, el Sistema de Evaluación del Desempeño Profesional Docente (en adelante, Evaluación Docente), también conocido como Docentemás, se implementa desde el año 2003 y constituye una evaluación obligatoria para los más de 70.000 docentes de aula que realizan su trabajo en establecimientos municipales. En este sistema, el Portafolio es el instrumento central utilizado para evaluar a los docentes, el cual recoge evidencia directa acerca de la enseñanza, la evaluación y las decisiones pedagógicas de los docentes. Incluye evidencia escrita y la filmación de una clase. Los portafolios son corregidos por profesores de aula seleccionados que cuentan con al menos 5 años de experiencia docente y se desempeñan en el mismo nivel, subsector y modalidad educativa que aquellos docentes a quienes evalúan. Estos correctores son capacitados por medio de rigurosos protocolos que buscan garantizar una corrección confiable.

Los portafolios docentes son ampliamente utilizados en Estados Unidos, ya sea para certificar la licenciatura en pedagogía, como es el caso del PACT (Performance Assessment for California Teachers), en la recertificación docente y en programas que acreditan excelencia pedagógica como el National Board Certification (Pecheone & Chung, 2006; Zeichner & Wray, 2000). En el contexto latinoamericano, Chile y Cuba son pioneros desde el año 2002 en la utilización del portafolio para evaluar competencias docentes (Sun, Calderón, Valerio y Torres, 2011; Valdés, 2006). Por su parte, Perú y Honduras han realizado implementaciones piloto de evaluación de sus profesores utilizando portafolios con miras hacia una evaluación nacional (Arnold y Feder, 2006; Ministerio de Educación de Perú, 2011).

El presente estudio se realizó para poder informar y mejorar los procesos de entrenamiento y monitoreo de la evaluación de portafolio escrito de la evaluación docente en Chile, y así aportar al mantenimiento y fortalecimiento de su validez. Para ello, la investigación se centró en la comprensión de la relación entre las representaciones mentales de los correctores de portafolio y la precisión evaluativa alcanzada por los mismos, utilizando la metodología de Pensamiento en Voz Alta (en adelante PVA; mejor conocida en inglés como Think Aloud).

En el transcurso de los dos últimos decenios, un limitado número de investigaciones se ha dedicado a estudiar lo que sucede en las mentes de los correctores mientras realizan su tarea (Bejar, 2012; Crisp, 2012). Dentro de ellas, un ámbito importante de interés ha sido comprender cómo las representaciones mentales utilizadas por los correctores influyen en sus juicios evaluativos (Myford, 2012). Como investigadores, pensamos que realizar investigaciones de esta índole es relevante para poder aportar a la validez de la evaluación docente; sobre todo cuando se supone, tal como lo hacemos nosotros, que las representaciones que los correctores realizan de la evidencia y de los criterios de evaluación son el reflejo de sus propias subjetividades y, por otra parte, que los mismos procesos de capacitación y entrenamiento pueden producir distintas comprensiones en cada corrector. Esta investigación pretende aportar al conocimiento de este tema, siendo, a nuestro entender, la primera investigación de su naturaleza realizada en el contexto chileno y latinoamericano.[1]

Representaciones mentales y conceptos en contextos de evaluación

Las representaciones mentales, objeto de esta investigación, pueden ser entendidas como estados mentales, tales como pensamientos, creencias, deseos, percepciones e imágenes sobre cosas, que cuentan con intencionalidad y pueden ser evaluadas en términos de la consistencia, adecuación, precisión y verdad (Stanford Encyclopedia of Philosophy, 2008). Según van der Shaaf et al. (2005), los correctores que difieren en aquellas representaciones cognitivas que aplican debieran también diferir en sus juicios. Para el caso específico de correctores de evidencias escritas, por ejemplo, la puntuación asignada a un trabajo depende de la comparación realizada entre la pauta de evaluación o una representación existente de cómo debiera ser un trabajo ideal y la representación generada sobre lo evaluado (Crisp, 2012; Lumley, 2002). Así, las representaciones sobre las que se basan las comparaciones resultan esenciales para llevar a cabo un juicio preciso.

De manera similar, se ha observado que en cualquier tipo de resolución de tareas con una meta definida, las personas necesitan generar una representación mental de la situación a solucionar y manipular activamente dicha representación. La habilidad para representar un problema de manera eficiente pareciera ser mediada en alto grado por el conocimiento que la persona haya alcanzado de los conceptos implicados en la tarea (Jonassen, 2000). Distintos tipos de conceptos utilizados para el pensamiento son relevantes en cuanto forman la base sobre la cual se construyen representaciones mentales que, para el caso de los portafolios, son el principal insumo para su evaluación (Jucks & Paus, 2012; van der Shaaf et al., 2005).

Por otra parte, la lectura de trabajos a ser evaluados requiere de construcción de significado e interpretación por parte del corrector (Crisp, 2012). Son estas unidades de significado que la mente construye sobre su entorno, que se originan tanto de la experiencia cotidiana como de instancias de educación formal, a lo que se le denomina *conceptos* (Vygotski, 1934, 1991). Para el caso específico de tareas de alta demanda cognitiva, como lo es la evaluación de portafolios (Crisp, 2008), parece razonable pensar que la operación con conceptos categoriales sea bastante frecuente, ya que estos permiten disminuir la complejidad que requiere el procesamiento de los variados elementos que configuran el mundo percibido (Rosch, 2002).

Representaciones y validez de evaluación

Como se mencionaba más arriba, aquellos procesos mentales empleados por los correctores al evaluar influyen en la validez de las inferencias que son capaces de hacer las diversas mediciones. Distintos estudios han mostrado variabilidad en la consistencia de los correctores y su grado de severidad al corregir. Es por ello que los razonamientos cognitivos que efectúan quienes corrigen han comenzado a formar parte de las investigaciones asociadas a la evaluación y medición educacional (Bejar, 2012).

En tanto, los estándares internacionales para la medición y evaluación requieren que, al momento de puntuar un test, se documenten todos los procedimientos realizados para garantizar la precisión del proceso de puntuación. Debe monitorearse la frecuencia en que ocurren errores y detectarse las fuentes

[1] Al realizar una búsqueda utilizando términos como "representaciones mentales", "correctores", "evaluadores" y "portafolio", tanto en español como inglés, en bases de datos de psicología, educación, ciencias sociales y multidisciplinares tales como PSYCINFO, ERIC, SCIELO y Google Académico, no se encontró ninguna investigación que indagara representaciones mentales de correctores en Latinoamérica.

de estos de manera de poder eliminarlas. Asimismo, cuando la puntuación de un test involucra juicios hechos por personas, debe monitorearse regularmente que los evaluadores se apeguen a lo esperado (American Education Research Association, American Psychological Association & National Council on Measurement in Education, 1999). En consecuencia, esta investigación busca conocer los procesos cognitivos que realizan los correctores para emitir sus juicios evaluativos, monitorear su apego a lo esperado y conocer fuentes de error con el fin de orientar los procesos de capacitación y entrenamiento y solventar estas debilidades.

A la hora de analizar la validez de una evaluación es posible tomar prestado de la tradición de medición cuantitativa dos conceptos de alto interés: *subrepresentación* y *varianza irrelevante* del constructo evaluado (Messick, 1995). La subrepresentación de un constructo sucede cuando no se incluyen facetas o dimensiones importantes del constructo en su evaluación, mientras que la inclusión de varianza irrelevante de un constructo sucede cuando se incluyen aspectos asociados con otros constructos o se aplican métodos que afectan la pesquisa del constructo (Messick, 1995). Así, por ejemplo, en su estudio sobre las evaluaciones de portafolio válidas y confiables, Heller, Sheingold y Myford (1998) encontraron que la omisión de procesos esenciales, la falta de consideración de criterios de evaluación relevantes y la inclusión de criterios externos a los considerados en la pauta de corrección ponían en riesgo la validez de la evaluación. Asimismo, otros aspectos que amenazan la validez incluyen las conductas del evaluador que demuestren una confusión sobre los criterios de evaluación, la evidencia evaluada o la manera en que se conjugan ambas, así como la intromisión de criterios que podríamos llamar *dinámicos*, donde la calidad de evidencias evaluadas previamente afecta al juicio de aquellas evaluadas posteriormente (Myford, 2012).

Otros criterios irrelevantes pueden derivarse de la interpretación de intenciones que un corrector realiza de la información presentada por el agente evaluado. Tal como señala Schwarz (1996), en cualquier situación comunicativa las personas suponen la existencia de relevancia e intencionalidad de la información comunicada. Así, la representación de lo evaluado es en parte una recreación de aquellas intenciones de significado originales atribuidas a quien se evalúa (Crisp, 2008, 2012).

Por otra parte, de acuerdo con el modelo de Associated System Theory (Citado en Schleicher & Day, 1998), orientado al estudio de las representaciones, dos grandes ejes son los que categorizan las representaciones mentales: eje concreto/abstracto y eje autorreferente/heterorreferente. En línea con esta teoría, los investigadores han encontrado que una evaluación de portafolio es más confiable cuando se alterna entre representaciones concretas enfocadas en el material evaluado y representaciones abstractas enfocadas en predecir la calidad del trabajo de la persona evaluada en contextos distintos al del portafolio (van der Schaaf et al., 2005). De manera similar, las evaluaciones resultan ser más precisas al basarse en criterios heterorreferentes, o externos, como lo son los criterios de una pauta de corrección, que en criterios autorreferentes, o basados en estándares personales (Schleicher & Day, 1998).

Considerando la relevancia de los niveles de abstracción de las representaciones, los conceptos sobre los cuales se construyen dichas representaciones y los niveles de adecuación de estas respecto al constructo evaluado, a continuación nos dedicaremos a analizar estas tres cualidades de las representaciones mentales y su relación con el nivel de precisión de los juicios evaluativos. De esta forma, las preguntas de investigación que guiaron nuestro diseño exploratorio y análisis fueron: ¿hay alguna relación entre la cantidad y cualidad de las representaciones involucradas en la corrección y la precisión de la corrección? Si hay alguna relación entre las representaciones y la precisión de la corrección, ¿cómo es esta relación?

Al igual que Crisp (2008) y Heller et al. (1998), creemos que mejorar el entendimiento que se tiene con respecto a los procesos y representaciones cognitivas implicados en la evaluación de desempeños tiene el potencial de suscitar cambios en los procesos de evaluación y entrenamiento para alcanzar evaluaciones más válidas y confiables. Tal como lo sugiere Bejar (2012), la cognición de los correctores es de gran valor a la hora de poder acuñar sugerencias sobre cómo entrenarlos mejor. La presente investigación se llevó a cabo sobre la base de apreciaciones como esta.

Método

Participantes

Se realizó un muestreo por conveniencia en que se contactó a seis profesores que cumplieran con los siguientes criterios de inclusión: (a) estar participando por primera vez como correctores de portafolio en el proceso de Evaluación Docente, con el fin de excluir el efecto que las experiencias anteriores como corrector podrían tener en sus representaciones y (b) estar corrigiendo evidencia de las asignaturas de Lenguaje y Matemática en los distintos niveles en los que estos se evalúan (primer ciclo, segundo ciclo y enseñanza media —los equivalentes chilenos a la educación primaria y secundaria—). Esto último, con el objetivo de representar a sectores centrales de enseñanza que, además, corresponden a más de la mitad de las evaluaciones que anualmente se realizan en el Sistema de Evaluación Docente.[2]

Todos los participantes del estudio llevaban dos semanas de entrenamiento en la corrección de portafolio al momento del estudio. La primera semana incluyó la lectura y discusión acerca del significado, foco y límites de los criterios a ser evaluados, así como la aplicación de dichos criterios para evaluar portafolios. La segunda semana consistió en un período de constante aplicación de lo aprendido a través de correcciones de diversos portafolios. Las características de los participantes pueden revisarse en la Tabla 1.

Tabla 1
Características de los participantes

Género	Edad	Nivel de evidencia que corrige	Sector de evidencia que corrige	Años de experiencia en aula
Femenino	28	Primer ciclo	Matemática	5
Femenino	47	Educación media	Matemática	22
Femenino	45	Segundo ciclo	Lenguaje	17
Femenino	35	Educación media	Lenguaje	10
Femenino	40	Segundo ciclo	Matemática	13
Femenino	30	Primer ciclo	Matemática	10

Para contactar a los profesores participantes, se elaboró una lista aleatoria con todos los correctores que cumplían con los criterios de inclusión. Luego, el entrevistador acudió a las salas de corrección y contactó al primer corrector de su lista para explicarle de qué se trataba el estudio y preguntarle si le interesaba participar. Si no aceptaba, debía contactar al segundo corrector de su lista repitiendo este procedimiento hasta lograr la aceptación de alguno.

Todos los profesores accedieron a participar voluntariamente y firmaron un consentimiento informado en que se detallaron los objetivos del estudio, las responsabilidades de la participación, su carácter voluntario, la confidencialidad de la información entregada, la ausencia de riesgos por participar y los beneficios del estudio.

Naturaleza de los datos del estudio: portafolio y rúbricas utilizadas

Es importante considerar que la cualidad de los datos generados y analizados en este estudio depende del tipo de portafolio y rúbricas a partir de las cuales los correctores crean sus representaciones mentales y del método utilizado para acceder a dichas representaciones.

En el portafolio escrito utilizado, los docentes reportan lo realizado clase a clase durante una unidad pedagógica de 8 horas, dan a conocer la principal evaluación utilizada para medir los aprendizajes de sus alumnos durante dicha unidad y contestan distintas preguntas de reflexión acerca de sus prácticas pedagógicas (Sun et al., 2011). Las rúbricas de evaluación incluidas en el estudio miden algunos criterios

[1] Dato obtenido a partir de análisis propios elaborados con las bases de datos proporcionadas por el Sistema Nacional de Evaluación Docente.

de enseñanza y aprendizaje pertenecientes a 5 de las 8 dimensiones evaluadas en el portafolio escrito (Sun et al., 2011). En la Tabla 2 se pueden ver más detalles sobre el tipo de evidencia y rúbricas analizadas por los correctores del estudio (Docentemás, 2010; Flotts y Abarzúa, 2011).

Tabla 2
Ámbitos de los criterios de evaluación aplicados por los correctores del estudio

Entrada de portafolio	Dimensión de evaluación	Preguntas o instrucciones del portafolio	Ámbitos de los criterios de evaluación
Unidad Pedagógica	Organización de los elementos de la unidad	Describa lo realizado en cada clase de su unidad pedagógica.	Capacidad de organizar las clases de la unidad en una secuencia pedagógica coherente.
	Análisis de las actividades de las clases	Describa dos aspectos de su unidad que considere como fortalezas. Señale por qué las considera fortalezas.	Capacidad de reconocer las fortalezas de una unidad pedagógica y cómo estas favorecen el aprendizaje de los alumnos y los aspectos que dificultan dicho aprendizaje.
Evaluación de la unidad pedagógica	Calidad de la evaluación de la unidad	Adjunte la evaluación y criterios de corrección correspondientes que utilizó para evaluar los aprendizajes alcanzados por sus alumnos durante la unidad pedagógica.	Capacidad del docente de diseñar evaluaciones que permitan determinar en sus estudiantes el logro de los aprendizajes de una unidad pedagógica. Capacidad de formular instrucciones e ítems claros y de diseñar pautas de corrección que identifiquen adecuadamente las respuestas o desempeños esperados frente a dichas instrucciones o ítems.
Reflexión a partir de los resultados de la evaluación	Retroalimentación a un alumno	Escoja a un alumno que haya demostrado aprendizajes logrados y no logrados y transcriba lo que le diría para retroalimentarle sobre los resultados que obtuvo en la evaluación.	Capacidad del profesor para orientar a los alumnos para mejorar sus aprendizajes a partir del análisis de los aspectos logrados y no logrados en la evaluación.
Reflexión Pedagógica	Reflexión pedagógica	Describa algunas dificultades que haya enfrentado alguna vez con respecto al clima de aprendizaje. ¿A qué atribuyó estas dificultades? Mencione las acciones que realizó para reponer un clima propicio para el aprendizaje.	Capacidad que tiene el docente para identificar los factores que dificultan la existencia de un clima propicio para el aprendizaje y plantear acciones para restablecerlo.

Nota. Fuente: Docentemás, 2010; Flotts & Abarzúa, 2011.

Método de recolección de información

Para recolectar la información se utilizó la metodología de PVA, que exige enfocarse en tareas desafiantes mientras se da expresión verbal a los pensamientos que entran en la atención (Ericsson & Fox, 2011). Esta metodología se considera óptima para indagar la manera en que los evaluadores identifican evidencia que se corresponde con criterios de evaluación, la manera en que interpretan la evidencia y la manera en que terminan por puntuar dicho criterio (Heller, Sheingold, & Myford, 1998). Así, esta metodología proporciona información verbal acerca del razonamiento durante la resolución de una tarea o problema (Ericsson & Simon, 1993; Fonteyn, Kuipers, & Grobe, 1993; Van Someren, Barnard, & Sandberg, 1994).

En este contexto, la consigna dada a los correctores fue: "A fin de que yo entienda lo que va pensando, le pido que corrija este portafolio poniéndole volumen a su pensamiento, es decir, contando lo que va leyendo, pensando, las dudas que le van surgiendo, etc., y explicando cada paso de su razonamiento lo mejor que pueda". Así, se buscó conocer cómo los correctores razonan al corregir la evidencia.

El gran beneficio de esta técnica es que vincula el proceso de pensamiento con percepciones concurrentes, permitiendo al investigador acceder a información disponible en la memoria de trabajo (Lundgrén-Laine & Salanterä, 2010), la cual refleja la información procesada por las personas en sus pensamientos y, por ende, las representaciones que forman parte de los mismos. Recopilar datos por medio de PVA y luego analizar sus protocolos verbales es una buena manera de investigar procesos de decisión complejos y sobrepuestos (Lundgrén-Laine & Salanterä, 2010), tal como son los procesos involucrados en la corrección de portafolios (Heller et al., 1998).

Aun cuando se ha criticado esta técnica de investigación en puntos como su confiabilidad y posibilidad de replicar el análisis por posibles problemas de interpretación de intenciones entre sujeto estudiado y codificador (Zanov & Davison, 2010) y por eventuales problemas de reactividad del pensamiento registrado a las instrucciones, interrupciones y presencia del entrevistador (Schooler, 2011), esta técnica sigue siendo ampliamente utilizada en investigación psicológica en cuanto entrega grandes cantidades de información sobre el pensamiento (Ericsson & Fox, 2011).

Procedimiento

Profesores y psicólogos del equipo de Instrumentos de Evaluación de Docentemás realizaron una entrevista de pensamiento en voz alta apoyándose en una guía que detallaba el procedimiento. Esta definía que cada corrector debía corregir 14 indicadores de portafolio escrito usando pautas de evaluación específicas para cada evidencia y el método de PVA.

El entrevistador podía ocasionalmente interrumpir al corrector para solicitarle que le explicara en mayor detalle su razonamiento. Además, si se mantenía en silencio por más de 10 segundos, se le repetía un breve recordatorio para enfatizar que continuara "poniéndole volumen a su pensamiento".

Todas las entrevistas se realizaron en un mismo día de forma que todos los correctores hubiesen contado con el mismo período de aprendizaje y entrenamiento en la corrección y el tiempo no fuese una variable influyente en los resultados. En este contexto, los portafolios revisados correspondían a los que a cada corrector, de acuerdo con su grupo de corrección, le tocaba corregir ese día en el marco de un entrenamiento grupal. Estos portafolios habían sido seleccionados por el especialista encargado de entrenar a los correctores de ese grupo, cerciorándose de que estuviera completo y de que los contenidos fueran diferentes de los revisados en sesiones anteriores.

Análisis de datos

Para el análisis de datos, se comenzó por desarrollar un libro de códigos inicial a partir de la literatura revisada referida a los procesos cognitivos involucrados en este tipo de tareas (Crisp, 2008; Heller et al., 1998; van der Schaaf et al., 2005). Luego, este libro se complementó con los razonamientos observados en revisiones de extractos de los protocolos. Así, se elaboró una versión preliminar buscando que reflejara

lo más posible los contenidos de los razonamientos, y luego se comenzó la revisión de los protocolos. Con base en los nuevos contenidos que aparecían en las revisiones, se fueron agregando nuevos códigos y quitando o modificando algunos otros que se habían incluido inicialmente, para finalmente constituir un libro de códigos que se apegara a los razonamientos observados con el cual se corrigió toda la evidencia. La construcción de este libro de códigos se basó en la teoría fundamentada (Strauss & Corbin, 2002).

La unidad de análisis usada en el proceso de codificación de los protocolos fue cada unidad de significado dentro de un momento mental determinado, que tenía un significado posible de interpretar por sí mismo. Se codificaron tres tipos de procesos cognitivos: representaciones, juicios evaluativos y metacogniciones. En otra oportunidad se presentaron resultados sobre metacogniciones (Torres, García, & Leyton, 2012). Aquí nos dedicaremos a presentar los resultados referidos a las representaciones cognitivas.

Las representaciones cognitivas se codificaron según su nivel de abstracción y su cercanía o lejanía con respecto a los significados de pauta y evidencia objetivados por los codificadores.

Por otra parte, se determinó la precisión de cada indicador evaluado por un codificador, considerando si era precisa o imprecisa sobre la base de un criterio externo dado por miembros del equipo encargado de elaborar y entrenar a los correctores en la aplicación de las rúbricas de evaluación. Ellos indicaban el nivel de desempeño final que debía asignarse a la evidencia del indicador de acuerdo con los criterios establecidos en la pauta de corrección (Insatisfactorio, Básico, Competente o Destacado), y si este se condecía en forma exacta con el asignado por el corrector, se consideraba esa evaluación precisa, o de lo contrario, imprecisa.

De acuerdo con la precisión encontrada en la corrección de los 14 indicadores, se seleccionaron 6 de ellos para analizar en el estudio considerando dos criterios: (a) que refirieran a distintos ámbitos de evaluación del portafolio, tal como se detalla en la Tabla 2, a fin de contar con una muestra representativa del instrumento, y (b) que contaran con un número similar de correcciones precisas e imprecisas, con el fin de poder realizar análisis y conclusiones en relación con la precisión de corrección, unidad de análisis del estudio. Así, se analizaron 17 correcciones precisas y 19 imprecisas. Para cada una de estas, la pauta de evaluación tenía entre dos y cuatro criterios a considerar, lo cual generó una muestra de 90 episodios de pensamiento a analizar.

Todas las entrevistas fueron grabadas y transcritas. Asimismo, todos los protocolos fueron codificados en forma independiente por dos investigadores y posteriormente consensuados e ingresados al software ATLAS TI para su análisis.

Resultados

Debido a que los razonamientos analizados y reportados a continuación son el producto de la interacción entre la información que proporcionan las rúbricas de corrección (pautas) y las entradas de portafolio (evidencias), las codificaciones de los razonamientos se realizaron considerando tanto la pauta como la evidencia y la interacción entre ambas para cada episodio evaluativo. Asimismo, cada evaluación o corrección que realiz**ó** el corrector se clasificó según fuera precisa o imprecisa, sobre la base del criterio externo antes mencionado.

Presencia total de representaciones, juicios y metacogniciones

Se identificó un total de 1838 actividades cognitivas, de las cuales la gran mayoría correspondió a representaciones (59%) y, en menor medida, a juicios (22%) y metacogniciones (19%). Al analizar la proporción de representaciones mentales de los correctores sobre el total de sus actividades cognitivas según el nivel de precisión que alcanzaban en sus juicios evaluativos, no se encontraron diferencias importantes entre evaluaciones precisas e imprecisas (59% y 60% respectivamente). Es por ello que nos dedicaremos a revisar los resultados que nos permitan saber si existen diferencias en precisión según la *cualidad* o *tipo* de representaciones que emplean los correctores.

Representaciones mentales de acuerdo al nivel de abstracción, los tipos de conceptos a la base y el nivel de precisión de las evaluaciones

Niveles de abstracción. Las representaciones cognitivas que empleaban los correctores al evaluar se clasificaron en tres niveles de abstracción: *concretas*, *mediana abstracción* y *alta abstracción.* En la Tabla 3 pueden verse definiciones y ejemplos de estas.

Tabla 3
Definición y ejemplos de las representaciones según nivel de abstracción

Código	Definición	Ejemplo
Representación concreta	Cuando lo que se observa en el razonamiento del corrector era una lectura o referencia textual a la información de la evidencia o la pauta que refleja una reproducción de la información procesada.	El corrector lee la planificación de clases enviada por el profesor en el portafolio escrito.
Representación de mediana abstracción	Cuando se observa un etiquetamiento de la información de la evidencia o la pauta, a través de conceptos que ayudan a simplificar la carga cognitiva del procesamiento de la información.	El corrector luego de leer la evidencia, señala: "Aquí el docente señala una fortaleza de su clase, al referir que los recursos fueron adecuados a los intereses de sus alumnos".
Representación de alta abstracción	Cuando las representaciones se desprenden de manera indirecta de la información que entrega la evidencia o la pauta, incluidos los conocimientos o experiencias previas, predisposiciones a la evaluación e interpretaciones de alto orden, entre otros.	El corrector señala: "Este es el indicador que menos me gusta corregir". El corrector señala: "Los docentes siempre obtienen un mal desempeño en este indicador".

Al analizar los niveles de abstracción de las representaciones encontramos, en términos generales, que tanto las correcciones precisas como las imprecisas muestran una alta presencia de representaciones concretas (50%), seguida de una importante presencia de representaciones de mediana abstracción (40%) y una baja presencia de representaciones de alta abstracción (10%). Ahora bien, aun cuando ambos niveles de precisión coinciden en la manera en que se distribuyen los niveles de abstracción de sus representaciones, a la hora de compararlas en mayor detalle se pueden observar ciertas diferencias de interés. Aun cuando tanto las correcciones precisas como las imprecisas presentan igual proporción de representaciones de mediano nivel de abstracción (40% y 41% respectivamente), estas difieren en cuanto al grado de presencia de representaciones concretas y de alta abstracción, presentando un tipo de tendencia inversa. Específicamente, las correcciones precisas muestran una mayor presencia de representaciones concretas (52%) y una menor presencia de representaciones de alta abstracción (7%), mientras que, por el contrario, las evaluaciones imprecisas muestran una mayor presencia de representaciones de alto nivel de abstracción (13%) y una menor presencia de concretas (47%), tal como puede verse en la Tabla 4.

Tabla 4
Cantidad de representaciones según nivel de abstracción y precisión de la evaluación

	Representaciones concretas	Representaciones de mediano nivel	Representaciones de alto nivel
Precisa	52%	40%	7%
Imprecisa	47%	41%	13%

La baja presencia total de representaciones abstractas daría cuenta de que todos los correctores muestran una tendencia a apegarse a lo expuesto en la evidencia y la pauta. Asimismo, las representaciones que mayoritariamente utilizan las personas al corregir corresponden, por una parte, a lectura textual, y por otra, al etiquetamiento de información por medio de conceptos claves suficientemente apegados a la información leída. Por otra parte, en las correcciones precisas las representaciones que utilizan se apegan más a la lectura o referencia textual, en comparación con las imprecisas, que se abstraen en mayor medida de la información de la tarea.

Tipos de conceptos bajo las representaciones. Para seguir profundizando el análisis de las representaciones, estas se analizaron de acuerdo con subcategorías relativas al tipo de conceptos sobre las cuales se generaban, tal como se muestran en la Figura 1.

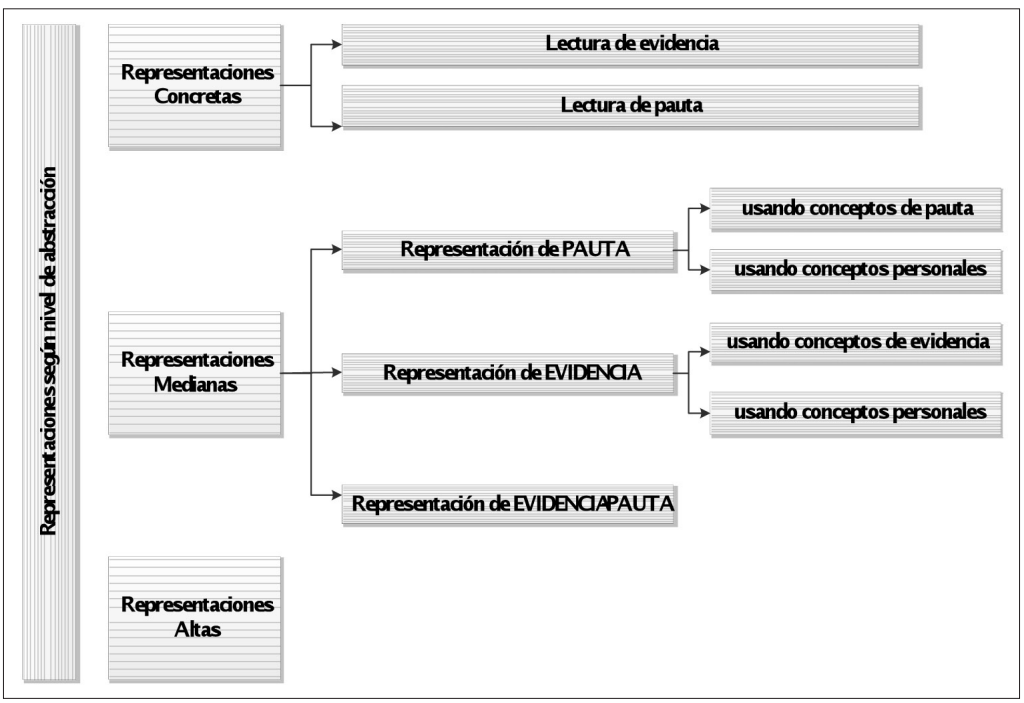

Figura 1. Esquema de tipos de representaciones concretas y de mediano nivel de abstracción.

En el caso de las representaciones concretas, estas se codificaron en dos subcategorías, a saber: *representaciones concretas de lectura de evidencia* y *representaciones concretas de lectura de pauta*.

En el caso de las representaciones de mediano nivel de abstracción, estas se subdividieron en tres subcategorías: *representaciones de pauta de mediana abstracción, representaciones de evidencia de mediana abstracción* y *representaciones de pauta-evidencia de mediana abstracción.*

Tabla 5
Definición y ejemplos de las subcategorías de representaciones de mediano nivel de abstracción

Subcategoría	Definición	Ejemplo
Representaciones concretas de lectura de evidencia	Cuando el corrector lee la información de la evidencia o realiza una referencia textual a la misma.	El corrector lee la evaluación de la unidad enviada por el profesor en el portafolio escrito.
Representaciones concretas de lectura de pauta	Cuando el corrector lee la información de la pauta de corrección o hace una referencia textual a la misma.	El corrector lee los criterios de evaluación del indicador que va a evaluar.
Representaciones de pauta de mediana abstracción	Cuando los razonamientos tienen por finalidad poder autoaclarar o simplificar criterios de pauta que se están utilizando por medio de conceptos claves o categorías. Estas representaciones, a su vez, pueden ser clasificadas con dos códigos distintos: *representaciones de pauta usando conceptos personales* y *representaciones de pauta usando conceptos de pauta*.	El corrector señala: "Para evaluar este indicador, es necesario revisar el hilo conductor en la planificación de la unidad".
Representaciones de evidencia de mediana abstracción	Cuando las representaciones están referidas a razonamientos que buscan simplificar la evidencia o comprenderla mejor. Estas representaciones, a su vez, pueden ser clasificadas con dos códigos distintos: *representaciones de evidencia usando conceptos personales* y *representaciones de evidencia usando conceptos de evidencia*.	El corrector señala: "Aquí el corrector está explicando cómo resolvió la dificultad que recién describió".
Representaciones de evidencia-pauta de mediana abstracción	Cuando las representaciones del corrector buscan evaluar la evidencia en función de la pauta, o bien comprender los criterios de pauta por medio de la consideración de hechos observados en la evidencia.	El corrector señala: "Aquí el docente explica lo que hizo para solucionar la dificultad que tuvo al desarrollar su clase…[según la pauta] esto es una decisión pedagógica".

Al analizar la precisión de las evaluaciones en función de las formas de representarse *la evidencia*, encontramos que para lograr evaluaciones más precisas, al abordarse la evidencia pareciera que es beneficioso representársela principalmente de manera concreta a través de la lectura (véase Tabla 6).

Tabla 6
Formas de representarse la evidencia y precisión de las evaluaciones

	Representación de evidencia			
	Concreta		Mediana abstracción	
	Lectura de evidencia	Representación evidencia/pauta	Representación de evidencia usando conceptos de evidencia	Representación de evidencia usando conceptos personales
Precisa	35%	17%	13%	2%
Imprecisa	24%	11%	13%	8%

Por otra parte, al analizar las formas de representarse *la pauta* en relación con la precisión de las evaluaciones, encontramos que para lograr evaluaciones más precisas pareciera que es más útil no representársela tanto de manera concreta, sino en un mediano nivel de abstracción en función de puntos de referencia relevantes (como la evidencia) u otras formas de etiquetamiento (véase Tabla 7).

Tabla 7
Formas de representarse la pauta y precisión de las evaluaciones

	Representación de pauta			
	Concreta		Mediana abstracción	
	Lectura de pauta	Representación evidencia/pauta	Representación de pauta usando conceptos de pauta	Representación de pauta usando conceptos personales
Precisa	17%	17%	3%	5%
Imprecisa	23%	11%	3%	3%

A la hora de comparar representaciones de mediano nivel de abstracción en evaluaciones precisas e imprecisas, es posible observar ciertas tendencias. Primero, los resultados muestran que tanto en evaluaciones precisas como imprecisas los correctores se representan *evidencia* utilizando *conceptos de evidencia* y la *pauta* utilizando *conceptos de pauta* en igual medida. Esto mostraría que categorizar la información de la pauta o de la evidencia con los conceptos presentes dentro de la misma pauta o evidencia no sería ni beneficioso ni perjudicial para la corrección. Ahora bien, cuando se representa la *pauta* por medio de conceptos personales se observa una tendencia positiva a favor de lograr evaluaciones más precisas, pero no así cuando se trata de representar la *evidencia* con dichos conceptos, llevando esto último a evaluaciones imprecisas.

Una última diferencia importante es que, cuando se es más preciso, se utiliza una representación conjunta de evidencia y pauta en mayor proporción que cuando se es impreciso, lo cual podría reforzar la idea de que internalizar la pauta y utilizarla como "lente" de corrección parece ayudar a lograr juicios más precisos. Los resultados anteriores se muestran en la Tabla 8.

Tabla 8
Naturaleza de conceptos con que se realizan las representaciones y precisión de la evaluación

	Etiquetamiento		Elaboración		
	Rep. de pauta usando conceptos de pauta	Rep. de evidencia usando conceptos de evidencia	Rep. de pauta usando conceptos personales	Rep. de evidencia usando conceptos personales	Rep. evidencia/ pauta
Precisa	3%	13%	5%	2%	17%
Imprecisa	3%	13%	3%	8%	11%

Representaciones mentales según nivel de adecuación y nivel de precisión de las evaluaciones

Un segundo nivel de análisis se relacionó con categorizar las representaciones de acuerdo con su nivel de adecuación, independientemente de su nivel de abstracción y de los tipos de conceptos utilizados. De acuerdo con esta categoría, las representaciones podían ser *adecuadas*, cuando se apegaban a la información que entregaba la evidencia o la pauta, o bien *inadecuadas*, cuando las interpretaciones del corrector se alejaban notoriamente de lo interpretado por los codificadores acerca de la pauta o la evidencia e inducían a error.

Se establecieron diversos subcódigos correspondientes a los distintos tipos de representaciones inadecuadas encontradas, las que, sobre la base de distinciones emergentes posteriores al análisis, se agruparon en dos categorías de acuerdo con el origen o causa que mostraba la inadecuación: por problemas en la interpretación o problemas en la selección de información. En la Tabla 9 puede verse la definición y los ejemplos de estas dos categorías.

Tabla 9
Definiciones y ejemplos de subcategorías de representaciones inadecuadas

Subcategoría	Definición	Ejemplo
Representaciones inadecuadas por problemas en la selección de información	Cuando los correctores incluyen información equivocada o menos información de la requerida en su evaluación.	El corrector confunde la parte de la evidencia que debe seleccionar para corregir un indicador. El corrector corrige un indicador con un criterio de pauta de otro indicador.
Representaciones inadecuadas por problemas en la interpretación de información	Cuando los correctores manifiestan problemas relacionados con la sobreinterpretación del significado de la información y/o la inclusión de criterios. personales al momento de evaluar la información. De esta forma, se observa que los correctores incluyen criterios idiosincráticos distintos de los presentes en la pauta de corrección al evaluar, o bien sobreinterpretan la información de la evidencia incluyendo interpretaciones propias sobre las intenciones que habría tenido el docente al realizar una determinada actividad.	El corrector señala: "Seguramente al docente no le funcionó la actividad que había planificado porque sus alumnos son muy desordenados, pero no se le puede castigar por eso". El corrector señala: "Si el docente tiene un error de ortografía aquí, no le puedo poner competente en este indicador, aunque cumpla con los criterios de la pauta".

Al analizar el nivel de adecuación del contenido representacional en relación con la precisión de las evaluaciones, se observa que la gran mayoría de los contenidos de las representaciones se adecúa a lo esperable, aun cuando las evaluaciones precisas presentan una menor proporción de representaciones inadecuadas (6%) en comparación con las imprecisas (10%), tal como puede verse en la Tabla 10.

Tabla 10
Nivel de apego del contenido representacional a la tarea y precisión de la evaluación

	Representaciones adecuadas	Representaciones inadecuadas
Precisa	94%	6%
Imprecisa	90%	10%

Al ahondar en este resultado analizando el origen o tipo de inadecuación de los contenidos representacionales, se encuentran diferencias importantes en cuanto al nivel de precisión de las evaluaciones. Las evaluaciones imprecisas presentan una fuerte presencia de representaciones inadecuadas por problemas con la *interpretación de información*. Las evaluaciones precisas presentan representaciones inadecuadas por problemas en la *selección de información*, ya sea por selección insuficiente o equívoca de esta, tal como puede observarse en la Tabla 11.

Tabla 11
Origen de inadecuación de los contenidos representacionales y precisión

	Representaciones inadecuadas	
	Problemas en la selección de información	Problemas en la interpretación de la información
Precisa	54%	46%
Imprecisa	29%	71%

De esta manera, podemos concluir que las representaciones inadecuadas constituyen un perjuicio a la corrección y afectan a su precisión. Asimismo, respecto de la causa u origen de las inadecuaciones de las representaciones, los resultados muestran que cuando los problemas de interpretación aumentan en forma importante, indicando una mayor inclusión de criterios externos a la evaluación en lugar de errores de selección de información, la validez de evaluación se ve más amenazada.

Discusión

El presente estudio tuvo por objetivo explorar posibles relaciones entre la cantidad y la cualidad de las representaciones mentales activadas en la corrección de portafolios docentes y la precisión de evaluación. La motivación para llevar a cabo esta exploración en el marco del Sistema de Evaluación del Desempeño Profesional Docente se basa en hallazgos generados en otros contextos de corrección que han determinado que las estrategias cognitivas y los procesos de evaluación que emplean los correctores son una importante fuente de varianza para la evaluación (Crisp, 2012; Myford, 2012). De esta forma, nuestro estudio esperaba realizar una doble contribución: aportar al conocimiento acumulado acerca de procesos mentales implicados en la corrección de portafolios y a la vez aportar al aseguramiento de la validez de corrección de los portafolios de la evaluación docente chilena.

Al revisar los resultados obtenidos en relación con los tipos de representaciones involucradas en la corrección de portafolios docentes, nos encontramos con que los evaluadores se representan la información en distintos niveles de abstracción (alto, medio y concreto), mostrando distinto nivel de adecuación o apego al significado convencional de las situaciones involucradas, pudiendo ser adecuadas o inadecuadas, y utilizando diferentes conceptos para representarse la información implicada (personales, de la pauta de evaluación y de la propia evidencia).

De acuerdo con nuestros hallazgos, en la corrección de portafolios los correctores utilizan mayormente representaciones concretas y medianas, predominando estas claramente por sobre las de alta abstracción. Estos resultados corroboran los hallazgos que han señalado que la operación con categorías es más frecuente cuando se llevan a cabo tareas de alta demanda cognitiva, ya que estas permiten disminuir la complejidad que requiere el procesamiento de elementos variados (Crisp, 2008; Rosch, 2002). Al ser una tarea de alta complejidad, la evaluación de portafolios estudiados demuestra una importante presencia de representaciones de mediana abstracción que utilizan categorías que etiquetan o elaboran la información revisada.

Ahora bien, con respecto a la relación entre la cantidad y la cualidad de las representaciones involucradas en la corrección y la eficacia de corrección, encontramos que efectivamente existe una relación entre las representaciones y la precisión evaluativa. Los resultados muestran que más que la cantidad de representaciones, lo que incide en la precisión de los juicios es la cualidad de estas.

De este modo, en general, representarse la información de evaluación de manera más concreta, así como por medio de ciertas formas de mediano nivel de abstracción, evitando representaciones altamente abstractas, pareciera ser una tendencia beneficiosa para obtener evaluaciones más precisas. En nuestros resultados, las representaciones de alto nivel de abstracción son más frecuentes en evaluaciones imprecisas que precisas.

Este resultado se diferencia en parte de los hallazgos de van der Shaaf et al. (2005), quienes han señalado que una evaluación de portafolio es más confiable cuando se alterna entre representaciones concretas y abstractas. Nuestros resultados mostrarían que más que la alternancia entre representaciones de alto y bajo nivel de abstracción, el nivel adecuado de abstracción de las representaciones parece depender de la naturaleza de los documentos revisados y de la tarea a realizar.

Así, nuestros resultados demuestran diferencias entre el tipo de representaciones que parecen ser beneficiosas a la hora de representarse la evidencia y la pauta. Al representarse la evidencia, encontramos que ayuda hacerlo de forma concreta, es decir, a partir de relecturas o alusiones textuales a esta, o bien mediante la representación conjunta con pauta. Por otra parte, al representarse la pauta de corrección, parece más conveniente hacerlo a un nivel de mediana abstracción, en función de conceptos personales o, como ya decíamos, representándosela al ir evaluando la evidencia misma.

De esta manera, los documentos descriptivos que intentan ilustrar hechos, como la evidencia, requieren de un nivel de abstracción más concreto en su representación, mientras que los documentos que se basan en conceptos técnicos abstractos y genéricos, como las rúbricas de evaluación, parecen requerir representaciones de un nivel de abstracción mayor. Dado que la naturaleza de la tarea de corrección requiere la aplicación de conceptos abstractos presentes en la rúbrica para analizar hechos concretos presentes en la evidencia, los niveles de abstracción medio parecen ser beneficiosos para relacionar ambos tipos de información.

Con respecto a estos resultados, es importante tomar en cuenta que aquel tipo de portafolio investigado por van der Schaaf et al. (2005) difiere de aquel utilizado en la Evaluación Docente en distintos aspectos que podrían explicar la diferencia en los hallazgos. Mientras que los correctores de la Evaluación Docente en Chile debían juzgar competencias docentes específicas por medio del análisis de evidencias o entradas escritas que se relacionaban uno a uno con cada rúbrica y sus variados criterios de evaluación, los correctores del estudio holandés eran evaluados de manera holística por medio de 8 escalas Likert a partir de entradas que además de incluir evidencias escritas, similares a las de la evaluación docente, incluían entrevistas a los profesores, videos de sus clases y evaluaciones que los alumnos hacían de ellos.

De esta manera, los correctores holandeses requerían considerar todos los tipos de evidencia para juzgar una cualidad específica del docente como un todo, lo cual implicaba una constante alternancia entre ideas generales abstractas y ejemplos concretos específicos al momento de decidir la competencia general del docente en alguna dimensión. Los correctores chilenos en tanto, debían considerar entradas concretas específicas para evaluar criterios de evaluación docente de un nivel de abstracción alto pero bastante menos abstracto que los holandeses.

Por otra parte, en cuanto a la adecuación de las representaciones, los resultados muestran que el uso de representaciones inadecuadas, alejadas de sentidos convencionales o de la tarea es poco frecuente en las evaluaciones. No obstante, al analizar este tipo de representaciones según el nivel de precisión de las evaluaciones, encontramos que las evaluaciones imprecisas muestran casi el doble de representaciones inadecuadas que las precisas.

Estas representaciones inadecuadas podían originarse por dos tipos de causas, a saber, dificultades para seleccionar la información requerida al corregir o bien dificultades en la interpretación de la información utilizada durante la evaluación. En las evaluaciones imprecisas, la inadecuación de las representaciones se relacionaría en mayor medida con problemas de interpretación de información, relacionados a su vez con la inclusión de criterios personales distintos de los indicados y con sobreinterpretaciones de la información evaluada, mientras que las evaluaciones precisas presentaban un mayor número de representaciones inadecuadas por problemas en la selección de información, tanto por selección insuficiente como por selección equívoca.

De esta forma, es posible afirmar que, cuando se incluyen criterios externos a la evaluación, tal como lo son las interpretaciones personales de conceptos claves de las pautas de evaluación (autorreferentes), existe un mayor riesgo de realizar evaluaciones menos precisas. Este hallazgo concuerda con lo planteado por Heller et al. (1998), quienes señalan que la falta de consideración de criterios de evaluación relevantes o la inclusión de criterios externos a los considerados en la pauta de corrección son una importante amenaza a la validez de la evaluación.

Por otra parte, en términos generales, nuestros resultados parecen ser coherentes con los hallazgos de Jonassen (2000), quien señala que la habilidad para representarse un problema de manera eficiente está mediada por el conocimiento que la persona haya alcanzado de los conceptos implicados en la tarea. Es bastante probable que dicha familiaridad se refleje en la capacidad de representarse la pauta de corrección por medio de conceptos propios y de representarse la evidencia en forma conjunta a la pauta, lo cual se observó en mayor medida en las evaluaciones precisas, e indica probablemente un grado mayor de internalización de los criterios de evaluación en comparación con quienes evaluaron de forma más imprecisa. Esta "traducción" o flexibilidad conceptual puede reflejar el nivel de manejo sobre un dominio abstracto específico. Por otra parte, la representación concreta de las prácticas y contenidos de clase observados en la evidencia podría ser el reflejo de la comprensión automática de escenarios pedagógicos familiares para los correctores docentes, quienes evalúan a profesores que enseñan en sus mismas disciplinas y niveles educativos.

Los hallazgos de esta investigación permiten orientar procesos de entrenamiento de correctores en procesos de evaluación. En esta línea, un primer aspecto a considerar tiene relación con el nivel de abstracción que se modela al entrenar a corregir. Los resultados mostrarían que cuando lo que se corrige es factual, como la evidencia entregada por los docentes, resulta positivo invitar a los correctores a representarse la información en un nivel concreto mediante relecturas o referencias textuales. En cambio, cuando la información es más abstracta, como lo son los criterios de evaluación, resultaría más conveniente ayudar a que los correctores elaboren la información y la expresen de manera adecuada en sus propios

términos. Al hacer esto se ayuda a los correctores a internalizar la pauta de manera significativa, lo cual les permite posteriormente evaluar de forma más precisa.

Otro aspecto que se desprende de los resultados tiene relación con los errores que aparecen al corregir. En esta línea, al entrenar es especialmente importante tener en consideración qué problemas en la interpretación afectan a la precisión de las evaluaciones, y es necesario trabajar sobre ellos al entrenar a los correctores, junto con apoyarlos para que seleccionen la información adecuadamente. Así, por ejemplo, pedir a cada corrector que comunique sus propias interpretaciones de conceptos claves de la pauta de corrección para llevarlas a discusión grupal entre distintos correctores en entrenamiento podría ser una buena forma de evitar problemas de interpretación que podrían afectar a varias evaluaciones. Por otra parte, la generación de documentos de apoyo para el evaluador, estructurados especialmente para el vaciado de información relevante para la corrección de ciertos indicadores que resultan más difíciles de corregir, podría ayudar a disminuir los errores en la selección de información.

No obstante, los desafíos para los sistemas de entrenamiento no solo se encuentran a nivel de aquello que los correctores aprenden durante el período de capacitación y práctica, sino que en aquellas creencias pedagógicas e identidades docentes que los correctores traen consigo a la evaluación. Tal como señala Richardson (1996), las creencias personales son utilizadas por los docentes para apoyar sus decisiones y llevar a cabo juicios de otras personas. Ciertas creencias, como lo son los valores personales, juegan un rol central en los sistemas de creencias e identidades de los docentes, son difíciles de cambiar y pasan a teñir percepciones e influir en los comportamientos docentes (Maxwell-Smith & Esses, 2012). A fin de reducir al mínimo el efecto que estas creencias pedagógicas pueden tener sobre las representaciones que los correctores hacen sobre los criterios de evaluación (como la interpretación o inclusión de criterios externos), es importante enfatizar en la exploración de las creencias de los correctores por medio de la discusión de interpretaciones de la pauta señalada más arriba. Generar cierto nivel de conflicto cognitivo que permita el desequilibrio y cambio de aquellas creencias docentes (Mason, 2003) que entran en conflicto con las creencias pedagógicas a la base de la evaluación docente es fundamental.

Los resultados de esta investigación resultan iluminadores para una mejor comprensión del proceso de corrección de portafolios escritos de la evaluación docente. Comprender la "caja negra" del proceso de corrección puede tener grandes implicancias para el fortalecimiento de la validez del sistema, en la medida en que se utilice dicha comprensión para diseñar entradas de portafolio, pautas de corrección y procesos de entrenamiento que permitan a los correctores comprender y aplicar los criterios evaluativos de manera cada vez más cercana a los lineamientos pedagógicos que sustentan la evaluación. Poner énfasis en aquello que piensan los correctores es crucial para lograr una mayor validez, pues finalmente son ellos los instrumentos de medición de las prácticas que leen en sus pares evaluados.

Todavía queda un amplio espacio de desarrollo en torno a este tema y en relación con estos portafolios. Creemos, por ejemplo, que sería interesante que futuros estudios comparen las representaciones mentales de evaluadores ejemplares con aquellas de evaluadores promedio o aquellos que presentan mayores problemas de corrección. Esto permitiría extraer conclusiones que podrían incidir no solo en los procesos de entrenamiento, sino también en posibles estrategias de selección de los mismos evaluadores. Otro aspecto que sería relevante incluir es el análisis de la manera en que interactúan las representaciones con otros procesos mentales, tales como procesos metacognitivos. De esta forma, se podría analizar si existen patrones dentro de los episodios evaluativos que pudieran orientar y explicar en mayor medida la precisión de los correctores.

Por último, sería interesante replicar este estudio incorporando las recomendaciones planteadas sobre la corrección de clases grabadas, la cual también es parte de la evidencia del portafolio que entregan los docentes en esta evaluación. Esto permitiría identificar similitudes y diferencias con la corrección de evidencia escrita e informar y mejorar futuros procesos de entrenamiento, aproximándolos a las particularidades y exigencias de la corrección de ambos tipos de evidencia.

Finalmente, resulta importante señalar algunas limitaciones del estudio. Lo primero tiene relación con el tipo de muestreo empleado, que se realizó por conveniencia favoreciendo un proceso fácil y eficiente pero que pudo no haber suministrado las fuentes más ricas de información, como sí podría haberlo hecho un muestreo intencionado, tal como lo sugieren Martín-Crespo y Salamanca (2007).

Una segunda limitación inherente al método tiene relación con levantar la información del razonamiento empleado por los correctores al corregir la evidencia mediante la metodología de PVA. Aun cuando esto se discutió antes, es relevante señalar que existen preocupaciones metodológicas en relación con su uso. Una de ellas se refiere a que las verbalizaciones pueden ser una representación incompleta del pensamiento (Crisp, 2012), debido a que el pensamiento no lingüístico ocurre más rápido que el habla. De esta manera, pedirle al sujeto que verbalice sus pensamientos podría enlentecer el procesamiento de información y llevar a omisiones en las verbalizaciones (Ericsson & Simon, 1993). Sumado a lo anterior, otra preocupación ha sido si la verbalización puede cambiar el proceso que se estudia.

No obstante, aunque se sugiere precaución en las interpretaciones que se desprendan de los datos, la mirada general es que en situaciones adecuadas, la indagación mediante esta metodología produce información importante que no es posible obtener mediante otros métodos (Hayes & Flower, 1980; Kobrin & Young, 2003; Lumley, 2005, citado en Crisp, 2012).

Otra limitación tiene relación con la cantidad de correctores considerados para la realización de este estudio, los que no lograron representar la totalidad de los perfiles de estos, lo que habría permitido la realización de análisis asociados a sus desempeños. Sin embargo, dado que este estudio tenía un carácter exploratorio, no se buscó representatividad de los correctores sino de evaluaciones precisas e imprecisas.

Una última limitación se refiere al análisis empleado, que se hizo mediante códigos y subcódigos de los distintos razonamientos observados, excluyendo una codificación de lo inobservado y que resultaba esperable de acuerdo con la información presente en la evidencia y la pauta de corrección. Esto podría haber favorecido una comprensión más acabada del tipo de razonamientos que conducen a errores de interpretación y potenciado el uso que se podría dar a los datos generados con el estudio.

Con todas sus limitaciones, los estudios como el presentado siguen siendo de vital importancia para garantizar la validez de las correcciones llevadas a cabo por personas. Los estudios como este nos permiten comprender en mayor profundidad el proceso de razonamiento mental de los evaluadores cuando corrigen de manera precisa e imprecisa. Mediante la comprensión de estos fenómenos es posible orientar procesos de entrenamiento de los evaluadores de manera mejor informada. Los resultados de este estudio, no obstante valiosos, deben interpretarse con cautela, sobre todo considerando que, tal como concluimos, los tipos de representaciones más beneficiosas dependen mucho de las características de la tarea de evaluación.

El artículo original fue recibido el 30 de noviembre de 2012
El artículo revisado fue recibido el 28 de enero de 2013
El artículo fue aceptado el 31 de enero de 2013

Referencias

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

Arnold, R. y Feder, F. (2006). *Manual de herramientas de evaluación interna y externa. [*Documento de trabajo]. Recuperado el 19 de enero de 2013 de http://www2.gtz.de/wbf/4tDx9kw63gma/Manual_de_Herramientas_de_Evaluacion_Externa_e_Interna.pdf

Bejar, I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2-9. doi: 10.1111/j.1745-3992.2012.00238.x

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education, 38*(2), 247–264. doi: 10.1080/03057640802063486

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice, 31*(3), 10-20. doi: 10.1111/j.1745-3992.2012.00239.x

Docentemás (2010). *Manual Portafolio 2010 para primer ciclo*. Recuperado el 26 de enero de 2011 de http://www.docentemas.cl/docentes_documentos.php

Ericsson, K. A., & Fox, M. C. (2011). Thinking aloud is not a form of introspection but a qualitatively different methodology: Reply to schooler. *Psychological Bulletin, 137*(2), 351–354. doi: 10.1037/a0022388

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis*: *Verbal reports as data (Revised Edition).* London: The MIT Press.

Flotts, P. y Abarzúa, A. (2011). El modelo de evaluación y los instrumentos. En J. Manzi, R. González e Y. Sun (Eds.), *La evaluación docente en Chile* (pp. 37-61). Santiago, Chile: Centro de Medición MIDE UC.

Fonteyn, M. E., Kuipers, B., & Grobe, S. (1993). A description of think aloud method and protocol analysis. *Qualitative Health Research, 3*(4), 430-441. doi: 10.1177/104973239300300403

Hayes, J. R., & Flower, L. (1980). Identifying organization of writing process. En L. Gregg & E. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Mahwah, NJ: Lawrence Erlbaum.

Heller, J. I., Sheingold, K., & Myford, C. M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment, 5*(1), 5-40. doi: 10.1207/s15326977ea0501_1

Jonassen, D. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*(4), 63-85. doi: 10.1007/BF02300500

Jucks, R., & Paus, E. (2012). What makes a word difficult? Insights into the mental representation of technical terms. *Metacognition and Learning, 7*(2), 91–111. doi: 10.1007/s11409-011-9084-6

Kobrin, J. L., & Young, J. W. (2003). The cognitive equivalence of reading comprehension test items via computerized and paper and pencil administration. *Applied Measurement in Education, 16*, 115-140. doi: 10.1207/S15324818AME1602_2

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*(2), 255-264. doi: 10.1037/0021-9010.86.2.255

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing 19*(3), 246-276. doi: 10.1191/0265532202lt230oa

Lundgrén-Laine, H., & Salanterä, S. (2010). Think-aloud technique and protocol analysis in clinical decision-making research. *Qualitative Health Research, 20*(4), 565-575. doi: 10.1177/1049732309354278

Martín-Crespo, M. C. y Salamanca, A. B. (2007). El muestreo en la investigación cualitativa. *Nure Investigación, 27*, 1-4.

Mason, L. (2003). Personal epistemologies and intentional conceptual change. En G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 199-236). Mahwah, NJ : Erlbaum.

Maxwell-Smith, M.A., & Esses, V.M. (2012). Assessing individual differences in the degree to which people are committed to following their beliefs. *Journal of Research in Personality, 46* (2), 195-209. doi : 10.1016/j.jrp.2012.01.009

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

Ministerio de Educación de Perú (2011). *Sistematización del plan piloto de evaluación del desempeño docente.* Lima: Autores.

Myford, C. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues and Practice, 31*(3), 48-49. doi: 10.1111/j.1745-3992.2012.00243.x

Pecheone, R., & Chung, R. (2006). Evidence in teacher education. The erformance Assessment for California Teachers (PACT). *Journal of Teacher Education, 57(1), 22-36.* doi: 10.1177/0022487105284045

Richardson, V. (1996). The role of attitudes and beliefs in learning to teach. En J. Sikula (Ed.), *Handbook of research on teacher education (2nd ed.)* (pp. 102-119). NY: Macmillan.

Rosch, E. (2002). Principles of categorization. En D. J. Levitin (Ed.), *Foundations of cognitive psychology: Core readings* (pp. 251-270). Londres: The MIT Press.

Schleicher, D. J., & Day, D. V. (1998). A cognitive evaluation of frame of reference rater training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73*, 76-101. doi: 10.1006/obhd.1998.2751

Schooler, J. W. (2011). Introspecting in the spirit of William James: Comment on Fox, Ericsson and Best (2011). *Psychological Bulletin, 137*(2), 345–350. doi: 10.1037/a0022390

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation.* Nueva Jersey: Lawrence Erlbaum.

Stanford Encyclopedia of Philosophy (2008). *Mental representation.* Recuperado el 25 de noviembre de 2012 de http://plato.stanford.edu/entries/mental-representation/

Strauss, A. y Corbin, J. (2002). *Bases de la investigación cualitativa: técnicas y procedimientos para desarrollar la teoría fundamentada*. Medellín: Universidad de Antioquía.

Sun, Y., Calderón, P., Valerio, N. y Torres, P. (2011). La implementación de la evaluación docente. En J. Manzi, R. González e Y. Sun (Eds.), *La evaluación docente en Chile* (pp. 65-89). Santiago: Centro de Medición MIDE UC.

Torres, P., García, M. R., & Leyton, C. (2012, septiembre). *Metacognition involved in the assessment of teacher Portfolios*. Trabajo presentado en la 5th Biennial Meeting of the European Association for Research on Learning and Instruction (EARLI), en Milán, Italia.

Valdés, H. (2006). *La evaluación del desempeño del docente: un pilar del sistema de evaluación de la calidad de la educación en Cuba.* Recuperado el 19 de enero de 2013 de http://www.docentemas.cl/dm06_experiencias_ant04.php

Van der Schaaf, M., Stokking, K., & Verloop, N. (2005). Cognitive representations in raters' assessment of teacher portfolios. *Studies in Educational Evaluation, 31*(1), 27-55. doi: 10.1016/j.stueduc.2005.02.005

Van Someren, M. W., Barnard, V.F., & Sandberg, J. A. (1994). *The think aloud method. A practical guide to modeling cognitive processes*. Londres: Academic Press.

Vygotski, L. S. (1934/1991). *Obras escogidas, Tomo II, Pensamiento y lenguaje*. Madrid: Visor/MEC.

Woehr, D., & Huffcutt, A. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*(3), 189-205. doi: 10.1111/j.2044-8325.1994.tb00562.x

Zanov, M. V., & Davison, G. C. (2010). A Conceptual and empirical review of 25 years of cognitive assessment using articulated thoughts in simulated situations (ATSS) think-aloud paradigm. *Cognitive Therapy and Research, 34*(3), 282-291. doi: 10.1007/s10608-009-9271-9

Zeichner, K., & Wray, S. (2000). The teaching portfolio in US teacher education programs: What we know and what we need to know. *Teaching and Teacher Education, 17*(5), 613–621. doi: 10.1016/S0742-051X(01)00017-8