FONDO DE INVESTIGACIÓN Y DESARROLLO EN EDUCACIÓN
DEPARTAMENTO DE ESTUDIOS Y DESARROLLO
DIVISIÓN DE PLANIFICACIÓN Y PRESUPUESTO
MINISTERIO DE EDUCACIÓN

INFORME FINAL

UN INSTRUMENTO ONLINE PARA EVALUAR

COMPETENCIAS EVALUATIVAS DE DOCENTES DE EDUCACIÓN BÁSICA

INSTITUCIÓN ADJUDICATARIA: PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
INVESTIGADORA PRINCIPAL: LORENA MECKES GERARD
EQUIPO DE INVESTIGACIÓN: CARLA FÖRSTER, MARIANELA NAVARRO, ENRIQUE INOSTROZA
PROYECTO FONIDE: FX11668

SANTIAGO, 2018

Secretaría Técnica FONIDE, Departamento de Estudios y Desarrollo. División de Planificación y Presupuesto, MINEDUC.
Alameda N.º 1371, Piso 8. E-mail: fonide@mineduc.cl/web: www.fonide.cl

MONTO ADJUDICADO: $ 42.225.133

NÚMERO DE DECRETO: 1564

FECHA DEL DECRETO: 01/12/2016

INCORPORACIÓN O NO DE ENFOQUE DE GÉNERO: No

TIPO DE METODOLOGÍA EMPLEADA: MIXTA

CONTRAPARTES DEL CENTRO DE ESTUDIOS: Paula Guardia (CE), María Angélica Mena (UCE)

ESTA PUBLICACIÓN ESTÁ DISPONIBLE EN WWW.FONIDE.CL

ÍNDICE

Resumen

Se desarrolló una propuesta de instrumento de medición online del nivel de competencias de profesores de primer ciclo básico en evaluación de aprendizajes. Este instrumento tiene por objetivo realizar o complementar un diagnóstico de necesidades de formación continua en este ámbito. En este instrumento se presentan tares para evaluar las habilidades de los docentes para: recoger evidencia sobre el aprendizaje, analizar e interpretar evidencias de aprendizaje, retroalimentar formativamente, y certificar o calificar el aprendizaje logrado. Esta propuesta de dimensiones a considerar y sus respectivos indicadores o manifestaciones se realizó a partir de la revisión documental, y de instrumentos disponibles para evaluar el desempeño docente y la alfabetización en evaluación de aprendizajes; la realización de sesiones de consulta a expertos y el análisis de evidencias disponibles de Docente Más y Asignación de Excelencia Pedagógica. El instrumento actualmente disponible online, consiste en 20 tareas de evaluación y un reporte automático que recibe el docente después de concluirlo. Este es el resultado de la selección de preguntas a partir de la aplicación piloto una muestra de 398 docentes que respondieron 42 ítems distribuidos en dos formas. Adicionalmente, como resultado del proyecto se obtuvo un cuestionario de autorreporte validado, también a través de la misma aplicación piloto, también disponible online, el que será de utilidad para estudiar las prácticas evaluativas de los docentes.

Palabras Claves: evaluación de aprendizajes-profesores de educación básica o primaria.

1. INTRODUCCIÓN

El documento que se presenta a continuación comienza con una contextualización de las competencias docentes de evaluación a partir de antecedentes empíricos y teóricos; es decir, cómo se entienden las competencias evaluativas de los docentes; cómo se han investigado; y cuáles son las principales conclusiones de estudios que contribuyen a elaborar la propuesta de interés. Por lo tanto, el objetivo de este apartado es caracterizar la situación actual respecto a los avances en el conocimiento sobre este tema, argumentar su relevancia en el contexto nacional (ej. para la formulación de políticas públicas) e introducir los objetivos del proyecto.

Posterior a la presentación del objetivo general y los objetivos específicos, se desarrolla el marco teórico basado en la revisión de literatura sobre las dimensiones que conforman el constructo a evaluar. Este marco sustenta la operacionalización de las dimensiones que se evalúan mediante los instrumentos.

A continuación del marco teórico, se describe la metodología que da cuenta de los pasos seguidos en el proceso de construcción de los instrumentos de evaluación y la fase de aplicación piloto. Además, se especifican los criterios de selección de los participantes de la fase de pilotaje, y la caracterización de la muestra. Por último se incluyen los análisis psicométricos realizados, y los reportes que se generan para informar los resultados.

Luego se presenta el capítulo de resultados que da cuenta de los datos que sustentan las decisiones tomadas para llegar a conformar los instrumentos. Las conclusiones se basan en la discusión de los resultados y su contrastación con la bibliografía revisada. Para terminar, se recoge las conclusiones y plantea recomendaciones para las políticas públicas.

2. ANTECEDENTES Y CONTEXTUALIZACIÓN

La evaluación de los aprendizajes de los estudiantes es una de las responsabilidades más críticas de los profesores (Mertler & Campbell, 2015) por lo que las competencias en evaluación de los docentes tienen un gran impacto en el aprendizaje de los estudiantes (Black, 2004; Black & Wiliam, 1998; Tejedor & García-Varcárcel, 2010; Torres & Cárdenas, 2010; Wiliam, Lee, Harrison, & Black, 2004), definiendo las expectativas que se tienen de los alumnos e incidiendo directamente en la promoción de curso y en las opciones de educación superior (Equipo de Tarea para la Revisión del SIMCE, 2015).

Si bien la evaluación de los aprendizajes en el aula es una preocupación entre los docentes y recientemente para la política pública, poco se discute en torno al tema (Torres & Cárdenas, 2010) y ha sido un aspecto escasamente observado en América Latina, tanto por los investigadores como por los actores educativos (Mercado & Martínez, 2014; Prieto & Contreras, 2008; Ravela, Leymonié, Viñas, & Haretche, 2014). En contraste, se ha prestado mayor atención a la medición externa (Goubeaud, 2010), particularmente en Chile, a través del Sistema de Medición de la Calidad de la Educación [SIMCE]: “La evaluación realizada por los profesores en las salas de clase ha sido, hasta ahora, invisible para el sistema nacional de evaluación de aprendizajes y desatendida por las políticas educativas”

(Equipo de Tarea para la Revisión del SIMCE, 2015, p.9). El estudio sobre formación inicial docente en evaluación educacional realizado por la Agencia de la Calidad de la Educación (2016) da cuenta de falencias importantes en la preparación de los docentes en este ámbito por parte de las universidades.

En este contexto, en 2017 el Ministerio de Educación ha presentado al Consejo Nacional de Educación, los Criterios y Normas Mínimas Nacionales sobre Evaluación, Calificación y Promoción Escolar de estudiantes de Educación Regular en sus niveles Básico y Medio formación General y Diferenciada. Asimismo, el Plan de Evaluaciones presentado por el Ministerio de Educación y aprobado por el Consejo Nacional de Educación también incluye, además del calendario de pruebas SIMCE e internacionales, el desarrollo de estrategias para fortalecer la evaluación de aula. Por este motivo, es necesario investigar sobre las competencias en evaluación, saber qué saben los profesores sobre evaluación de aprendizajes y, sobre todo, cómo toman decisiones al evaluar el aprendizaje de sus estudiantes. Esto último hace necesario disponer de instrumentos válidos y confiables que apunten a medir las competencias evaluativas de los docentes y no solo su dominio conceptual sobre evaluación, que es lo que se incluye en la mayoría de las pruebas que disponibles (Gotch & French, 2014).

Investigar sobre las competencias en evaluación es importante pues remite a lo que los docentes enseñan y a los aprendizajes que los profesores promueven y valoran (Ravela, 2009; Vinas-Forcade & Emery, 2015), pues "las tareas que los maestros proponen a sus alumnos para evaluar el aprendizaje constituyen uno de los mejores indicadores del curriculum implementado" (Ravela, 2009, p. 56).

2.1. Importancia de la evaluación en el aprendizaje

La evaluación es un componente central del proceso de enseñanza - aprendizaje (Goubeaud, 2010) y una competencia esencial de todo profesor (Castillo & Cabrerizo, 2003). En efecto, existe una necesidad de fortalecer la evaluación de aula por su relación con el aprendizaje que logran los estudiantes; un docente que realiza buenas prácticas de evaluación promueve mayores aprendizajes significativos en sus estudiantes y por lo tanto mejores resultados (Tejedor & García-Varcárcel, 2010; Torres & Cárdenas, 2010).

Según las expectativas planteadas en el Marco para la Buena Enseñanza (Mineduc, 2008) y en los estándares orientadores de la formación inicial docente (Mineduc, 2012), se espera que el profesor comprenda la evaluación como un proceso sistemático de obtención de evidencias, en función del cual tome decisiones para mejorar su enseñanza y el aprendizaje de sus estudiantes. Desde esta perspectiva, la evaluación en ningún caso es una actividad meramente técnica, "sino que constituye un elemento clave en la calidad de los aprendizajes, condicionando la profundidad y el nivel de los mismos" (Villardón, 2006, p. 58).

Otro importante referente respecto de la competencia evaluativa de los docentes es el Marco para la Enseñanza de Danielson (2013). En este se considera que una buena práctica es tal si hay coherencia entre lo que se evalúa y los objetivos o expectativas de aprendizaje, la complejidad de los aprendizajes a lograr y cómo se evalúa. El marco contempla también la capacidad del docente de proponer diversas estrategias de evaluación para ofrecer distintas oportunidades a los estudiantes de demostrar lo aprendido.

2.2. Las prácticas evaluativas de los docentes

Si bien la evaluación se considera una dimensión sustantiva de todo proceso educativo, las investigaciones muestran que las prácticas evaluativas de los docentes, más que apoyar el aprendizaje lo obstaculizan (Prieto & Contreras, 2008). En efecto, en el aula predomina un enfoque tradicional donde la evaluación se concibe como fuera del proceso pedagógico y con un enfoque mayoritariamente memorístico e instrumental (Celman, 2005; Goubeaud, 2010; Organisation for Economic Co-operation and Development, [OECD], 2005; Prieto & Contreras, 2008; Sanmartí, 2007). De este modo, los docentes brindan excesiva preponderancia a la medición y a la certificación, lo que para los estudiantes se traduce en una mayor importancia por la calificación que por lo que están aprendiendo; asimismo, los profesores utilizan la evaluación como un instrumento de control, especialmente en lo que se refiere a normar conductas disruptivas. A este respecto, Torres y Cárdenas (2010) señalan que "continúan las prácticas de evaluación que buscan calificar, amenazar, sancionar o simplemente cumplir con un requerimiento institucional" (p. 149). En cuanto a los criterios de evaluación, muchas veces no son conocidos por los estudiantes y cuando estos existen, tampoco se considera a los alumnos en su definición, pues todo lo referente a la evaluación se concibe como un trabajo exclusivo del docente. Esto último tiene como consecuencia que el estudiante no tenga oportunidades para reflexionar sobre su propio desempeño, corregir sus errores y regular su aprendizaje (Sanmartí, 2007). Además, el error no se considera como una oportunidad para aprender sino como una conducta no deseable (Torres & Cárdenas, 2010).

Asimismo, los docentes ponen el énfasis en la evaluación de contenidos conceptuales, privilegiando la reproducción del conocimiento y prestando poca atención a la evaluación de habilidades cognitivas superiores (Prieto & Cárdenas, 2008). En cuanto a las prácticas de retroalimentación del aprendizaje, en general, son de carácter valorativo, con poca descripción del desempeño del estudiante, escasa orientación y oportunidades para reflexionar sobre lo realizado y las dificultades encontradas (Ravela, 2009). En contraste, los profesores que diseñan evaluaciones de mayor calidad potencian el desarrollo de habilidades cognitivas más complejas al mismo tiempo que pueden mejorar la motivación de sus estudiantes (Jensen, McDaniel, Woodard, & Kummer, 2014). Asimismo, la evaluación formativa cuando está bien planificada y ejecutada es una de las prácticas pedagógicas con mayor impacto en el aprendizaje que se traduce en la mejora de las estrategias de enseñanza y en retroalimentación de calidad para los estudiantes (Black, 2004; Black & William, 1998). Otras evidencias muestran una buena retroalimentación puede aumentar en 32 puntos los resultados de aprendizaje en lectura y matemática de SIMCE (Agencia de Calidad de la Educación, 2015a).

De acuerdo con Black (2004) se han identificado tres problemas principales en la práctica evaluativa de los aprendizajes: 1) los métodos de evaluación utilizados no son efectivos en la promoción del aprendizaje, 2) las prácticas de calificación ponen énfasis en la competencia entre los estudiantes, más que en el progreso individual y 3) un impacto negativo de la retroalimentación, especialmente en estudiantes de bajo desempeño, quienes se convencen de que no tienen capacidad y que no pueden aprender. Asimismo, en América Latina se detectaron también tres problemas en las prácticas de evaluación de los docentes de enseñanza básica. Estas pueden vincularse a las planteadas por Black (2004) y corresponden a: 1) el bajo nivel de demanda cognitiva de las tareas de evaluación que

proponen los docentes, 2) ausencia de ciclos de retroalimentación y 3) ausencia de criterios y procedimientos claros de calificación de los estudiantes (Ravela et al., 2014).

Específicamente en Chile, se encontró que un gran porcentaje de las evaluaciones contenían exclusivamente preguntas cerradas: 79% de las preguntas incluidas en pruebas de certificación son de selección múltiple, a lo que se agrega un 8% de preguntas cerradas de otros tipos. En tanto, el porcentaje de preguntas abiertas pasa de 13% en las pruebas certificativas a un 49% en las evaluaciones formativas. Esto último envía un mensaje contradictorio a los estudiantes: "lo importante es aprender a responder las preguntas cerradas, que son las que cuentan" (Ravela et al., 2014, p.26). Asimismo, las preguntas de los profesores chilenos presentan poca o nula información adicional o que planteen algún problema a resolver, en efecto, la proporción de preguntas que solo requieren recordar un concepto es la más alta en Chile (65%). Según los autores, este tipo de tareas promueve una retroalimentación centrada en la revisión de errores y no permite un mayor análisis o reflexión sobre aprendizajes más complejos. En cuanto a la calificación, el instrumento más valorado en Chile es la prueba escrita.

Otro antecedente importante lo aporta la evaluación docente obligatoria, donde se ha constatado que los profesores no logran buenos resultados en su desempeño al evaluar el aprendizaje. Dimensiones tales como "calidad de la evaluación" y "reflexión a partir de los resultados de la evaluación" no solo resultan ser de las más descendidas (Mineduc, 2015), sino las que de manera consistente, año tras año, resultan ser las más deficientes. A través del tiempo los profesores han mejorado su desempeño en la evaluación docente, no obstante, las dimensiones relativas a la evaluación son las que menos progreso muestran, lo que refleja que evaluar el aprendizaje de los estudiantes no es una habilidad fácil de adquirir ni de mejorar (Sun, Correa, Zapata y Carrasco, 2011).

Si bien las prácticas evaluativas de los docentes presentan muchas falencias, lo que es más preocupante es la falta de coherencia entre el discurso de los profesores en cuanto a sus concepciones sobre evaluación y sus prácticas (Ravela et al., 2014; Vinas-Forcade & Emery, 2015). Por ejemplo, para los docentes es importante desarrollar en sus estudiantes el pensamiento crítico, sin embargo, las tareas de evaluación carecen de contexto y relevancia y su resolución apela a habilidades cognitivas extremadamente simples (Ravela, 2009). En Chile se observó que los docentes asocian fuertemente los conceptos de *prueba – evaluación – calificación – nota*, pero en el discurso conciben la evaluación como una oportunidad de valorar a las personas, el aprender y el saber, que permite retroalimentar al estudiante sobre su aprendizaje y al profesor acerca de su quehacer pedagógico.

En efecto, mientras los docentes perciben sus prácticas evaluativas como cualitativas, formativas y constructivas, los estudiantes declaran que sus profesores muestran una excesiva preocupación por lo cuantitativo y lo sumativo, donde la finalidad de la evaluación es medirlos y acreditarlos (Torres & Cárdenas, 2010).

En síntesis, investigar sobre la competencia evaluativa de los docentes es importante pues la evaluación es el pilar desde donde se transforman las prácticas pedagógicas que impactan en el aprendizaje de los estudiantes (Black, 2004; Black & William, 1998; Tejedor & García-Varcárcel, 2010; Torres & Cárdenas, 2010). Pese a su relevancia, en Chile hay escasa investigación sobre la competencia evaluativa de los docentes, pues hasta ahora, se

le ha otorgado mayor importancia a la evaluación externa, mientras que ha habido un apoyo débil desde el Mineduc a la evaluación de aula o evaluación interna. Por ello, en el contexto de la Política de fortalecimiento de la evaluación de los aprendizajes que realizan los docentes, es indispensable indagar en las competencias de evaluación utilizadas en el aula (Equipo de Tarea para la Revisión del SIMCE, 2014).

2.3. Relevancia de la evaluación en aula para las Políticas Públicas

Si bien la evaluación de desempeño docente incluye en su portafolio de evidencias, procedimientos que permiten apreciar el desempeño de los docentes en este ámbito, este se aplica de modo esporádico (cada 4 años). Se trata de una evaluación de desempeño compleja que además de la evaluación de aprendizajes, tiene múltiples otros focos de atención. Por esta razón, contar con un instrumento ágil, aplicable a todos los docentes y con la capacidad de identificar fortalezas y debilidades específicas de evaluación, contribuiría a una política de fortalecimiento de la evaluación de aprendizajes en: (a) el diagnóstico fino de las necesidades de formación de las y los profesores en servicio, (b) el monitoreo de la eficacia de dicha política, y (c) la evaluación de la efectividad de las instancias de formación que se implementen. También sería beneficioso para el desarrollo de la investigación sobre las capacidades docentes en el ámbito de la evaluación de aprendizajes, pues los resultados que reporta actualmente la evaluación de desempeño en el ámbito de las prácticas evaluativas resultan aún muy generales para orientar decisiones específicas de formación y excluyen algunas áreas relevantes, como por ejemplo la retroalimentación a los estudiantes, la calificación o el análisis y uso de las evaluaciones externas.

Por otra parte, para desarrollar una Política de fortalecimiento de la evaluación de aprendizajes en el aula es necesario contar con un marco que defina las competencias de evaluación de los profesores, y describa su progresión. El presente proyecto propone que dicha descripción de niveles sucesivos de desarrollo se realice combinando el juicio experto con la evidencia empírica, es decir, definiciones teóricas y a priori de lo que constituirá un alto o bajo nivel de maestría, con el ordenamiento jerárquico que resulte de los patrones de respuesta de profesores a las tareas que les proponga el instrumento. Contar con una descripción de la progresión de la competencia evaluativa de los profesores desde niveles iniciales hasta niveles de experto(a), será de utilidad para que las instituciones de formación inicial cuenten con una descripción del nivel que se esperaría alcancen sus egresados en esta competencia, para que los propios docentes contrasten a lo largo de su carrera sus habilidades y conocimientos con estas descripciones, y para definir las oportunidades de formación que sean más pertinentes a su nivel.

3. PREGUNTAS DE INVESTIGACIÓN

Dadas las características de este estudio, no corresponde plantear preguntas de investigación.

4. HIPOTESIS U OBJETIVOS

Dadas las características de este estudio, no corresponde plantear hipótesis sino objetivos del estudio.

4.1. Objetivo general

Diseñar y validar instrumentos de diagnóstico online de las competencias y prácticas de los profesores de primer ciclo básico al evaluar el aprendizaje de los estudiantes.

4.2 Objetivos específicos

1. Diseñar instrumentos de auto aplicación online para diagnosticar el nivel de competencia en evaluación de aprendizajes que presentan los docentes que se desempeñan en enseñanza básica.
2. Realizar una aplicación piloto de los instrumentos para determinar sus características psicométricas.
3. Describir niveles de logro de la competencia evaluativa de los docentes a partir de los resultados obtenidos en la aplicación piloto.

5. MARCO TEÓRICO O CONCEPTUAL

La evaluación de los aprendizajes de los estudiantes para tomar decisiones pedagógicas que los ayuden a avanzar hacia la meta esperada es una de las competencias menos desarrollada en los docentes y que tiene un rol crucial en la forma en que los estudiantes aprenden (De Luca et al., 2015). En la bibliografía de habla inglesa se ha definido como alfabetización evaluativa (*assessment literacy*), aludiendo a la habilidad del docente para realizar evaluaciones confiables y asignar puntajes y notas que faciliten decisiones pedagógicas válidas coherentes con los lineamientos curriculares (Popham 2004, 2013; Stiggins 2002, 2004).

A partir de la revisión bibliográfica se han establecido cuatro dimensiones que conforman el constructo de competencia evaluativa de un docente y que constituyen las habilidades claves que un docente debe tener para evaluar el aprendizaje de sus estudiantes.

5.1. Recoger evidencia sobre el aprendizaje

El aprendizaje de los estudiantes es la meta final de la enseñanza en el aula y para establecer si esta meta se está logrando, es necesario recoger información de dicho proceso que nos permita emitir un juicio. Esta dimensión aborda dos ámbitos en los que un docente debe tener competencia: la recogida de evidencias en aula asociada a la interacción pedagógica constante y el dominio en la selección y elaboración de instrumentos y situaciones evaluativas pertinentes al contexto, las características de los estudiantes y al objetivo de aprendizaje que se desea evaluar.

Existen diferentes formas de categorizar las fuentes de evidencia del aprendizaje de los estudiantes según sea la forma en que lo expresa y quién levanta la información. Desde lo que se expresa, Griffin (2007) señala que las posibles fuentes de evidencia sobre el aprendizaje de un estudiante son: lo que este dice, escribe, o hace. Al respecto, Harlen (2007) señala que si bien los productos elaborados por los estudiantes (sus representaciones, informes, etc.) pueden ser una fuente rica de evidencia, la observación del proceso y de su toma de decisiones, aporta más. Por ejemplo, no es lo mismo analizar la hipótesis planteada en un informe de laboratorio de ciencias, que observar o escuchar el proceso de discusión del grupo que llevó a su formulación.

Otra forma de categorizar las fuentes de evidencia es de acuerdo con quién la levanta: es posible que surja a propósito de la interacción entre profesor y alumno, pero también puede ser posible que la interacción entre estudiantes permita generar información sobre el aprendizaje. En cualquier caso, el profesor es quien debe diseñar múltiples formas en que el estado actual del aprendizaje del alumno pueda hacerse evidente. Wiliam (2011) listó 50 formas de levantar evidencia sobre el razonamiento de los estudiantes, entre ellas, pedir que elaboren preguntas de pruebas con su respuesta correcta, o que clasifiquen producciones hechas por alumnos de años anteriores, argumentando respecto del orden propuesto.

También, las estrategias de recogida de evidencias de aprendizaje pueden presentar particularidades según asignatura. Por ejemplo, en la interacción profesor-alumno en una clase de Ciencias, esto puede darse a través de las preguntas que realiza el profesor, las cuales gatillan la expresión de hipótesis o explicaciones intuitivas de los estudiantes frente a un fenómeno, para así poder involucrarlos en un proceso de razonamiento productivo; y luego el docente puede invitar a varios de ellos a manifestar sus respuestas y perspectivas para, finalmente, ofrecer comentarios o nuevas preguntas ante sus respuestas (Chin, 2007). En Matemática en cambio, ante el planteamiento de una tarea, ejercicio o problema, el docente puede formular preguntas destinadas a que el estudiante explique su razonamiento, para luego cuestionarlo o sugerir nuevas rutas. Mientras que en el aprendizaje de la lectura, las preguntas que formula el docente frecuentemente estarán orientadas a indagar sobre la comprensión del texto por parte de los estudiantes.

También es importante que el docente sea capaz de distinguir la intencionalidad o propósito de la evaluación que realiza ya que las decisiones y selección de situaciones evaluativas será diferente si se trata de una evaluación formativa o sumativa (Shepard, 2006). En el caso de la elaboración de una propuesta de evaluación, es esencial hacer un adecuado diseño o bien realizar un análisis crítico de los instrumentos disponibles para seleccionar aquellos que resultan más adecuados para los propósitos evaluativos (Covacevich, 2014).

En cualquier caso, resulta clave que la evaluación cualquiera sea su ámbito (interacciones pedagógicas o situaciones evaluativas más estructuradas) cumplan con unos criterios mínimos que aseguren la calidad de la evidencia que se recoge. Así, la validez de contenido, es decir la coherencia entre la tareas o preguntas diseñadas o seleccionadas por el docente y el objetivo de aprendizaje sobre los que se quiere obtener evidencia (Brualdi, 1999; García, 2002; Hogan, 2004; Lukas y Santiago, 2004), su concordancia con las oportunidades de aprendizaje otorgadas a los estudiantes y la suficiencia de información que dichas tareas permiten generar, son criterios de calidad del desempeño docente en este ámbito (Förster y Rojas-Barahona, 2008).

Asimismo forman parte, tanto la capacidad del docente para elaborar o seleccionar tareas que son relevantes o nucleares para el aprendizaje que se ha desarrollado o que está en desarrollo, y que demandan poner en juego habilidades de orden superior, como su capacidad para seleccionar o diseñar tareas que permitan distinguir diferentes niveles de comprensión o de desarrollo de un aprendizaje (Hutchinson, Francis y Griffin, 2014). Por otra parte, Heritage y Heritage (2011) plantea que resulta crítico el grado en que las situaciones que plantean los docentes favorecen -o no- que los estudiantes hagan explícito su razonamiento y su nivel de comprensión, y no solo busquen verificar si sus respuestas son correctas o incorrectas. Esto es relevante ya que acceder al razonamiento del estudiante permite al docente tomar

mejores decisiones sobre la estrategia pedagógica a seguir y qué retroalimentación entregar para potenciar su aprendizaje.

La formulación o selección de tareas, preguntas, instrucciones e instrumentos para calificar o hacer una apreciación del desempeño de sus estudiantes, de modo que estas efectivamente permitan obtener información válida y confiable, también forma parte de esta dimensión. Se ha visto que el conocimiento que tienen los docentes respecto de cómo se elabora un instrumento de evaluación no es suficiente para predecir una buena práctica evaluativa, pues disocian su conocimiento teórico de la aplicación en su práctica cotidiana (Deneen & Brown, 2016; Popham, 2011).

Esta dimensión es relevante y ha sido abordada en distintos estándares y descripción de niveles de desempeño tanto de formación de profesores como de docentes en ejercicio de distintos países (Danielson, 2013; Joint Committee on Standars for Education Evaluation (2003); Kahl, Hofman, Bryant, 2013; Ministerio de Educación, 2012). En estos documentos se plantea que los docentes deben ser competentes en el desarrollo y elección de métodos de evaluación y en los procesos de recopilación de información en las evaluaciones, es decir, los métodos de evaluación deberían ser apropiados y compatibles con el propósito y contexto de la evaluación y los estudiantes deben tener diversas y suficientes oportunidades para demostrar los conocimientos, habilidades, actitudes o conductas que se les están evaluando. Estos lineamientos son la base de esta dimensión.

5.2. Analizar e interpretar evidencias de aprendizaje

Esta dimensión alude a la interpretación de muestras de aprendizaje de los estudiantes y da cuenta de la capacidad de los docentes para comprender la información que recogen e interpretar los resultados de las evaluaciones en el contexto específico de la asignatura y las características de sus estudiantes. Estas interpretaciones deben corresponder a representaciones precisas e informativas del rendimiento de un alumno en relación con las metas y objetivos de aprendizaje evaluadas (Joint Committee on Standards for Educational Evaluation, 2003).

El análisis de evidencias individuales supone la capacidad del docente para interpretar la información obtenida durante el proceso de aprendizaje de un estudiante, tanto de los productos elaborados (sus representaciones, informes, etc.) como de la observación del proceso de aprendizaje (por ejemplo, no es lo mismo analizar un cuento escrito por un estudiante que observar el proceso de elaboración de dicho texto). Pero el análisis en sí mismo no es lo que hará que un estudiante avance en su aprendizaje, la clave está en la pertinencia de las decisiones pedagógicas que el docente tome a partir de dicha evidencia.

De acuerdo con Dewey (1928), esta habilidad de interpretación requiere una capacidad de observación mucho más aguzada que la que se necesita para analizar los resultados de un test. Para ello, el docente debe tener dominio del conocimiento pedagógico del contenido de la disciplina que enseña (Shulman, 2005) pues esto le permite identificar las dificultades que está teniendo un estudiante en su razonamiento mientras se enfrenta a una tarea y dar las orientaciones necesarias (Zohar y Schwartzer, 2005). Tanto la habilidad para interpretar las respuestas y producciones del estudiante, como la que se requiere para formular o seleccionar tareas que permitan obtener información valiosa respecto de su aprendizaje,

están fuertemente influidas por la comprensión del docente respecto de la disciplina en que se inscribe su práctica evaluativa, como por su conocimiento de cómo esta se aprende. A su vez, para el análisis y uso de información, es clave que el docente cuente con una estructura que le permita contrastar la información y emitir un juicio a partir de la evidencia recogida. Este referente puede estar dado por la comprensión que tiene el docente sobre cómo progresa el aprendizaje del área disciplinaria que está enseñando o por la definición de una progresión de aprendizaje establecida explícitamente y que le permite contrastar la evidencia con este 'mapa de progreso' (Black, Wilson & Yao, 2011). Las progresiones del aprendizaje son comprensiones gradualmente más sofisticadas de conceptos y principios centrales o descripciones típicas de la trayectoria de un aprendizaje, en un dominio determinado dada una enseñanza apropiada. Ofrecen un marco interpretativo que permitiría entender la evidencia y situarla en una progresión de aprendizaje, pero además permiten orientar la toma de evidencia que es estratégico recoger. Contar con una comprensión sobre como progresa el aprendizaje en la disciplina que se enseña supone que el docente domina las formas en que se aprenden los ámbitos específicos de ella para, en base a este contraste, poder tomar decisiones para lograr los aprendizajes (Hutchinson, Francis & Griffin, 2014). Por este motivo, los instrumentos que se proponen en este proyecto se anclan en contextos disciplinares específicos de las asignaturas; pues hay evidencia suficiente que señala que un profesor con un buen dominio del contenido pedagógico es capaz de interactuar con sus estudiantes potenciado preguntas durante su clase que le permiten levantar información específica y de mayor riqueza sobre la forma en que están razonando (Carr et al., 2000; Jones & Moreland, 2005; Mishra & Koehler, 2006).

Los docentes efectivos interpretan y utilizan cotidianamente la información de sus evaluaciones para tomar decisiones sobre su enseñanza y así lograr mejoras en el aprendizaje de sus estudiantes (Bambrick-Santoyo, 2010). Sin embargo, la formación inicial de profesores se tiende a concentrar en el uso de datos de evaluaciones basadas en pruebas, los cuales están frecuentemente desconectados de la enseñanza o de la mejora de la escuela (Bocala & Boudett, 2015). Además, los profesores tienen creencias variadas sobre el uso de datos; por una parte, algunos sienten que carecen de la capacidad para utilizarlos y no los ven como un insumo para retroalimentar su práctica docente. Por otra parte, los docentes sienten que el uso de datos es una exigencia "externa" más asociada a la rendición de cuentas que al uso pedagógico que le pueden dar en el aula y no le ven utilidad, por tanto, no tienen la necesidad de usar la información (Ingram, Louis, & Schroeder, 2004). En esto el contexto escolar no ayuda, los colegios cada vez se centran más en la medición de aprendizajes, aplicando pruebas y contratando servicios externos lo que se traduce en más carga para los profesores que no saben cómo abordar estos informes e integrar los resultados a su quehacer diario.

Se ha visto que la capacidad de los profesores para usar información de evaluaciones y sus creencias sobre el uso de los datos se moldean dentro de sus comunidades profesionales, en las sesiones de capacitación y en sus interacciones con los otros docentes más experimentados, directores y consultores (Datnow & Hubbard, 2016). Además, la capacidad del docente para analizar datos incluye no solo su habilidad para interpretar información explícita en una tabla o informe de resultados, sino también analizar información implícita (por ejemplo la que surge de comparar datos), formular hipótesis para explicar resultados específicos, establecer la validez que tienen las

comparaciones de resultados que se hacen cotidianamente para distintos cursos o años, interpretar adecuadamente el promedio de un estudiante o de cursos en relación con un dominio de aprendizaje, y la integración de información de distintas fuentes, entre otros (Mandinach & Gummer, 2012) para tomar decisiones pedagógicas.

El tipo y complejidad de los análisis que los docentes realizan con información típica de procesos evaluativos escolares (resultados de una prueba, informe de notas de un curso, resultados de pruebas de nivel, informes de pruebas nacionales, de la corporación o red, ranking, número de notas, entre otros), y la toma de decisiones pedagógicas a partir de la interpretación de dicha información son habilidades relevantes que completan el ciclo evaluativo.

En síntesis, la dimensión de análisis e interpretación de la información supone la capacidad del docente para empatizar cognitivamente con su estudiante y formular hipótesis respecto de lo que está pensando o sobre el nivel de comprensión que ha alcanzado al dar una respuesta o ejecutar una tarea específica. También es relevante que el profesor comprenda, observe y analice la evidencia en relación con el aprendizaje que busca promover, y ubique con precisión a un estudiante o al curso en un nivel de progreso en relación con una producción o un conjunto de evidencias. Por último, debe ser capaz de analizar datos cualitativos y cuantitativos comunes en la práctica diaria de los profesores.

5.3. Retroalimentar formativamente

La retroalimentación es un elemento esencial de la evaluación para el aprendizaje y uno de los factores que mayor impacto tiene en los estudiantes (Hattie & Timperley, 2007; Hattie, 2009), ya que a través de ella puede aportar al alumno información sobre su desempeño y favorecer su comprensión, el desarrollo de sus habilidades y conocimientos, y promover su reflexión o el pensamiento crítico. Por ello, la literatura utiliza el concepto de retroalimentación formativa -en inglés formative feedback- (Shute, 2008). Brookhart (2008) destaca tres elementos claves de una retroalimentación para que esta cumpla una función formativa: (a) proporcionar información sobre la práctica pedagógica, (b) orientar la toma de decisiones sobre la enseñanza, basadas en la información recogida, y (c) proporcionar andamiaje al estudiante respecto de cómo mejorar su desempeño. Nuestro foco se sitúa en este tercer elemento.

Ahora bien, la retroalimentación puede ser oral o escrita, en sala de clases suele ser oral y centrarse en las interacciones o intercambios dialógicos, que ocurren de manera continua en el aula, en tiempo real, todos los días (Leahy, Lyon, Thompson y Wiliam, 2005; Ruiz-Primo, 2011). En el diálogo el alumno no solo recibe información desde el profesor, sino que se produce un intercambio de ideas, lo cual es esencial para que la retroalimentación sea efectiva (Laurillard, 2002). En un diálogo, el comentario de un estudiante puede reflejar su comprensión incompleta o imprecisa de un concepto, y desencadenar un evento de retroalimentación. La respuesta del profesor ante esta evidencia es generalmente rápida, espontánea y flexible, puede tomar diferentes formas, como la verificación de la precisión de la respuesta, la explicación de la respuesta correcta, consejos y ejemplos prácticos (Shute, 2008) o bien, puede responder con una pregunta, pedir al estudiante que explique su comentario, solicitar a otros estudiantes expresar su opinión o realizar una demostración (Ruiz- Primo, 2011).

La retroalimentación escrita, en tanto, no se da en tiempo real, el plazo para la interpretación de la evidencia puede ser más amplio, su uso más limitado y su carácter más formal respecto de la retroalimentación que ocurre durante el desarrollo de una clase. La retroalimentación escrita puede clasificarse según su forma o naturaleza (Ruiz-Primo y Li, 2013).

- Según su forma, puede ser: (a) un puntaje, porcentaje o nota; (b) un visto bueno o una marca que indica que se logró la respuesta esperada; (c) comentarios, palabras o frases; (d) rúbricas que proporcionan información evaluativa y/o criterios de puntuación; (e) o ser ilegible, es decir, proporcionar información que no se comprende.
- Según su naturaleza, se tiene: (a) evaluación sobre la completitud, esto es trabajo completo o incompleto; (b) evaluación sobre la calidad, la cual señala el nivel de comprensión, de imprecisiones o errores del trabajo, o fortalezas sin explicación (ej. ¡buen trabajo!); (c) edición, en que se edita o se modela cómo se debe hacer; (d) descripción, en la que se describe por qué está correcto o incorrecto; (e) exploración, en que se sondea el pensamiento del estudiante, por ejemplo a través de preguntas (ej. ¿siempre?); (f) transición, la cual indica la necesidad de una comunicación verbal.

En general, la retroalimentación del aprendizaje se ve de manera unilateral, es decir, desde el profesor al alumno; sin embargo, esta comunicación puede incluir a los estudiantes como participantes activos del proceso (Boud y Molloy, 2013; Hattie & Timperley, 2007). Más aun, se recomienda incluir la evaluación de pares como estrategia para que el feedback sea factible y más profundo sin sobrecargar al docente (Race et al., 2005; Panadero y Brown, 2017; Topping, 1998).

Puede haber distintas formas en que los estudiantes se involucren en la retroalimentación formativa, ya sea a través de la autoevaluación o de la evaluación de pares. En este proceso el docente es un guía; sin embargo, el nivel de orientación que proporciona puede ser desde bajo a alto. Por ejemplo, un nivel bajo es solicitar a los estudiantes que se autoevalúen o que se entreguen retroalimentación entre ellos sin criterios explícitos, y un nivel más alto, sería proporcionar a los estudiantes una rúbrica con criterios de desempeño y con ejemplos específicos del producto final o de cada uno de los niveles de logro, para utilizar este material al retroalimentarse recíprocamente (Panadero, Alonso- Tapia y Huertas, 2012).

En suma, entenderemos la retroalimentación del aprendizaje como la información comunicada al estudiante en respuesta a lo que él dice, a lo que él hace, a lo que él escribe o frente a cualquier otro tipo de desempeño o tarea (Hattie y Timperley, 2007; Hattie, 2009). Lo que hemos incluido en el ámbito de la retroalimentación del aprendizaje comprende a la habilidad del docente para actuar frente al desempeño de un estudiante. De acuerdo con Tunstall y Gipps (1996), la retroalimentación puede ser evaluativa, entendida como enjuiciadora (sancionando o premiando, aprobando o desaprobando el trabajo o respuesta del estudiante), o descriptiva (entregando información para mejorar o bien la reflexión del estudiante para promover un progreso del aprendizaje). Nos importa también la oportunidad que ofrece el docente para que los estudiantes participen de la retroalimentación y el nivel de orientación que les entrega para este fin (Panadero et al., 2012). Estas distintas formas de retroalimentar tienen diverso valor formativo o

incidencia en el aprendizaje, en esta categorización, la retroalimentación descriptiva tiene mayor impacto en el aprendizaje de los estudiantes (Kluger & De Nisi, 2000).

5.4. Certificar o calificar el aprendizaje logrado

La práctica de certificar el aprendizaje logrado consiste en sintetizar el logro alcanzado por los estudiantes a través de símbolos, que pueden ser letras, números, términos o breves descripciones sobre su desempeño, con el propósito de que esta información sea comunicada a terceros (ej. padres, colegas, otros profesionales, Ministerio de Educación). Una de las formas más comunes para certificar el logro es la calificación (entendida como la nota obtenida por un estudiante); sin embargo, esta no es la única forma de representar los aprendizajes obtenidos. El juicio profesional docente sobre la evidencia del aprendizaje puede manifestarse a través de una descripción basada en el análisis e interpretación de los logros alcanzados por los estudiantes (Zlokovich, 2001).

A diferencia de lo que ocurre con la evaluación formativa, la investigación sobre la calificación señala que es una de las formas más comunes para certificar el logro, pero está escasamente desarrollada y existe menos evidencia sobre las prácticas de calificación que mejor se relacionan con aprendizaje (Brookhart, 2012). Los focos de la investigación hasta ahora han estado en la frecuente falta de claridad y transparencia en los criterios para juzgar el aprendizaje logrado (Zlokovich, 2001) y en describir en qué medida las calificaciones resultan confiables y válidas, especialmente por la tendencia que existe a juzgar el aprendizaje con información limitada o insuficiente (Gimeno, 2010), o por incluir en las notas el esfuerzo, la percepción de habilidad y el comportamiento junto con el logro del aprendizaje[1] (Cross y Frary, 1999, McMillan, 2001, Randall y Engelhart, 2010).

Entre las prácticas de calificación que recomienda la bibliografía especializada, están: calificar el logro del aprendizaje que se está evaluando sin 'contaminar' dicha calificación con factores ajenos, como esfuerzo o puntualidad; no calificar las evaluaciones diagnósticas o formativas; comunicar previamente los criterios; y usar mecanismos para incrementar la confiabilidad. Pese a la relevancia que tiene la certificación del logro para la promoción escolar, lo interesante es que las prácticas de los docentes en este ámbito parecen ser bastante 'resistentes' a programas de formación continua (McMunn, Schenck & McColskey, 2003), probablemente porque las prácticas recomendadas por los especialistas en evaluación colisionan con valoraciones muy arraigadas, y porque las notas cumplen múltiples funciones en la vida escolar, más allá de representar una síntesis válida y confiable del logro de los aprendizajes alcanzados. En consecuencia, existe poca base para establecer a priori criterios de progresión en este ámbito del quehacer evaluativo del docente. Por ello, subrayaremos el carácter hipotético y tentativo de esta dimensión, especialmente si consideramos que es muy probable que la formación y las prácticas habituales de los docentes se distancien de estos ideales recomendados ya sea porque no han tenido oportunidad de aprender sobre ellos o porque en la práctica profesional sus decisiones deben atender y sopesar múltiples consideraciones respecto de las cuales la 'teoría evaluativa' resulta limitada.

[1] Aunque de todas formas, de acuerdo con esta línea de investigación, el factor que más pesa en las calificaciones asignadas por los docentes es el logro.

5.5. La evaluación de aprendizajes como parte del conocimiento pedagógico del contenido

El conocimiento pedagógico del contenido (CPC) incluye las mejores formas de representar las ideas de la disciplina, las analogías, ilustraciones, demostraciones, explicaciones o ejemplos más poderosos; en otras palabras, las representaciones más útiles para hacer el contenido compresible para otros. El conocimiento pedagógico del contenido incluye también una comprensión de aquello que hace que determinados tópicos sean más fáciles o difíciles de aprender, las concepciones y preconceptos que los estudiantes de distintas edades y contextos traen consigo al aprender los tópicos más frecuentemente enseñados (Shulman, 1987). Esta definición clásica del CPC no explicita que las prácticas de evaluación de aprendizaje forman parte de este conocimiento pedagógico del contenido. Sin embargo, más tarde esta dimensión se ha estudiado e incluido como componente constitutivo del CPC. Es así como en el modelo de Magnusson, Krajcik y Borko, (1999) se incluye el conocimiento de la evaluación de aprendizajes en Ciencias como uno de los componentes constitutivos del conocimiento pedagógico del contenido de los profesores de Ciencias. También se ha definido como competencia general de un docente (Jones y Moreland, 2005; Shulman, 2005).

De acuerdo con Earl (2003), para una evaluación apropiada del aprendizaje, los docentes necesitan recurrir a su conocimiento de los estudiantes y su comprensión del currículum y del contexto en que se da la evaluación. Por ello, la práctica evaluativa depende de la integración de diferentes dominios de conocimiento en un proceso, que determina la eficacia del profesor en dicha práctica. Si bien puede haber conocimientos genéricos y básicos referidos a evaluación de aprendizajes, que inciden en la capacidad del docente para lograr un buen desempeño en este ámbito (como saber que dar a conocer los criterios de evaluación anticipadamente es deseable, o que proporcionar retroalimentación descriptiva resulta más útil que solo aprobar o reprobar la producción de un estudiante), esto no es posible sin una adecuada comprensión de la disciplina, y de los objetivos de aprendizaje a alcanzar en ella.

Considerando lo anterior, concebimos las habilidades evaluativas como parte del conocimiento pedagógico del contenido de los docentes, y por lo tanto, lejos de intentar 'despejar' o neutralizar la incidencia del conocimiento disciplinar requerido para realizar una evaluación de aprendizajes válida y relevante, hemos optado por definir que el conocimiento disciplinar y pedagógico de las disciplinas que son enseñadas por un docente de educación básica, forman parte de dicha habilidad para evaluar aprendizajes.

En síntesis, la capacidad para recoger evidencias de aprendizajes tanto en la interacción pedagógica oral como en a través de productos escritos, las habilidades para analizar e interpretar evidencias de aprendizaje, la calidad de la retroalimentación que el docente da para que sea efectivamente formativa y la capacidad para certificar y calificar con precisión y validez los aprendizajes de los estudiantes, considerando que hay distinciones específicas asociadas a cómo se aprende cada disciplina, constituyen los elementos claves que se deben evaluar si se busca emitir un juicio respecto de las competencias evaluativas de los docentes.

6. METODOLOGÍA

Este proyecto es metodológico y consistió en elaborar dos instrumentos para evaluar conocimientos y prácticas de evaluación en aula. En esta sección se expondrá la metodología organizada de acuerdo con las etapas de desarrollo del proyecto, que se diagraman a continuación. Algunas de estas etapas no se desarrollaron del modo previsto (descripción de niveles de logro) debido a las características psicométricas del instrumento resultante. En esta sección se describen las etapas 1 y 2, mientras que la etapa 3, en su mayor parte, se describe en el apartado de Resultados.

6.1. Etapa 1: Definición del marco de evaluación y construcción de los instrumentos piloto

Esta etapa da cuenta de la consecución del objetivo específico 1: diseñar instrumentos de autoaplicación online para diagnosticar el nivel de competencia en evaluación de aprendizajes que presentan los docentes que se desempeñan en enseñanza básica. Las acciones seguidas para realizarlo fueron:

<u>Definición de las dimensiones a evaluar</u>

Esta acción se realizó a partir de la revisión bibliográfica y la opinión de expertos. Se definió con mayor precisión el concepto de competencia evaluativa de los docentes que se desempeñan en el sistema escolar y el propósito y usos esperados de los instrumentos. Entenderemos por competencia para evaluar el aprendizaje, como la capacidad del profesor para formular o seleccionar tareas que permitan recoger evidencias relevantes del aprendizaje de los estudiantes, interpretar sus producciones, intervenciones o respuestas en tanto

manifestaciones del nivel de aprendizaje o comprensión que ellos presentan; entregarles una retroalimentación con potencial para promover su aprendizaje, y tomar decisiones para dar cuenta a terceros de forma válida del nivel de logro alcanzado por ellos, todo ello en el contexto de las disciplinas que enseña. Esta clarificación permitió definir las dimensiones a evaluar y las características de los instrumentos (por ejemplo, el tipo de ítems y su uso).

Respecto del primer punto, para efectos de este proyecto, se generó una propuesta de dimensiones de la competencia evaluativa a partir de la revisión de distintos referentes entre los que están el Marco para la buena enseñanza (Mineduc, 2008), los Estándares orientadores de carreras de pedagogía (Mineduc, 2011, 2012), marcos internacionales (Danielson, 2013), las dimensiones abordadas por la evaluación docente y los lineamientos de evaluación definidos en las bases curriculares. Además, se realizaron 2 sesiones de consulta a expertos (ver detalle en Anexo Consulta a expertos ya reportado en Informe previo), los cuales eran especialistas en evaluación de aprendizajes, formación inicial y continua de profesores, con experiencia a nivel escolar, ministerial y académico. También se analizaron evidencias de Docente Más y Asignación de Excelencia Pedagógica compuestas por videos e instrumentos propuestos por los docentes y su reflexión a partir de ellos (ver detalle en Anexo Análisis ED y AEP ya reportados en Informe de septiembre de 2107).

También se definieron criterios de progresión para cada dimensión a partir de las cuales se elaboraron indicadores o manifestaciones jerarquizadas por el nivel de habilidad para cada dimensión, las que permitieron orientar la elaboración de las tareas o ítems.

La propuesta de constructo a evaluar se compone de 4 dimensiones y potenciales escalas que componen la competencia, tal como se sintetiza en la siguiente tabla:

Tabla 1. Dimensiones de la competencia evaluativa

Dimensión	Definición
1. Recoger evidencia del aprendizaje	Contempla el diseño de procedimientos de evaluación, la selección de tareas, su calidad técnica, demanda cognitiva y coherencia con el aprendizaje que se busca evaluar, así como el resguardo de principios éticos asociados. Considera el proceso a través del cual el docente obtiene información sobre el nivel de comprensión, o sobre el grado en que los estudiantes han desarrollado una habilidad. Incluye las habilidades docentes para diseñar situaciones en las que los estudiantes demuestren su desempeño o hagan explícita su comprensión, por ejemplo, formular preguntas adecuadas, observar sistemáticamente cómo los estudiantes resuelven un problema o desarrollan una tarea.

2. Analizar e interpretar evidencia del aprendizaje	Se refiere a la capacidad del docente para comprender e interpretar las evidencias de aprendizaje de sus estudiantes (intervenciones verbales, respuestas, desempeños) en relación con el aprendizaje y concluir a partir de datos cuantitativos provenientes de las evaluaciones aplicadas. Esta habilidad resulta clave para usar la evidencia como información para tomar decisiones pedagógicas que le permitan lograr las metas deseadas. Un docente preparado para leer e interpretar con exactitud la información que recoge a través de distintas formas de evaluación, podrá saber dónde están sus estudiantes en su aprendizaje, y será capaz de establecer metas más precisas y planificar una enseñanza eficaz para obtener los resultados esperados. Esto implica analizar evidencias tanto individuales como grupales, las cuales pueden ser de distinta naturaleza y origen, lo que requiere un profesor preparado para interpretar información cualitativa y cuantitativa.
3. Retroalimentar formativamente	Corresponde a la devolución y comunicación a los estudiantes de los resultados de la evaluación y al monitoreo del aprendizaje durante el proceso de enseñanza a través de la interacción entre profesor y estudiantes. Entenderemos la retroalimentación del aprendizaje como la información comunicada al estudiante o reacción de docente en respuesta a lo que el estudiante dice, a lo que él hace, a lo que él escribe o frente a cualquier otro tipo de desempeño o tarea (Hattie & Timperley, 2007; Hattie, 2009). Esta retroalimentación es formativa cuando trasciende la mera entrega de información y es utilizada para promover el aprendizaje. Considera también la oportunidad que ofrece el docente para que los estudiantes participen de la retroalimentación y el nivel de orientación que les entrega para este fin.
4. Certificar o calificar el aprendizaje alcanzado	Se refiere a las decisiones involucradas al asignar un símbolo (número, letra, breve descripción) para sintetizar el nivel de logro alcanzado por un estudiante y poder comunicarlo a ellos y a terceros. Considera también la capacidad del docente para reconocer prácticas de certificación que no son deseables y las consecuencias que esto tiene.

Definición de criterios e indicadores de progresión

A partir de la revisión bibliográfica para cada una de las dimensiones, se definieron criterios de progresión que orientaran la construcción de ítems.

1. Los criterios de progresión propuestos para recoger evidencia sobre el aprendizaje son los siguientes:

(a) ***El grado en que las tareas, estímulos, preguntas e instrumentos diseñados o seleccionados por el docente, presentan validez de contenido.*** Dicha validez se refiere a la correspondencia que existe entre el contenido/habilidades que evalúa el instrumento o tarea solicitada, y el campo de conocimiento al cual se atribuye dicho contenido (Brualdi, 1999; García, 2002; Hogan, 2004; Lukas y Santiago, 2004). Por lo tanto, se evidencia en el grado en que los instrumentos, preguntas, instrucciones, listas de chequeo o rúbricas que utiliza el docente, son coherentes con la información que desea recoger. Igualmente, abarca la consistencia entre tareas solicitadas y sus rúbricas de valoración. En este criterio se incluirá también el grado de relevancia en que el docente selecciona y formula tareas que resultan nucleares para el aprendizaje que se está desarrollando y el grado en que demandan que el estudiante ponga en juego habilidades complejas. Esto porque la validez de contenido no solo se traduce

en un alineamiento entre objetivos y tareas de evaluación, sino en que estas se dirijan a aspectos centrales del constructo y no a detalles menos relevantes (Förster y Rojas- Barahona, 2008).

Al respecto, la hipótesis es que el grado de dominio de los docentes se manifestará en la capacidad de discriminar claramente las tareas que son más coherentes con aquello que se busca evaluar, que van a lo nuclear y que favorecen o exigen que los estudiantes demuestren habilidades más complejas a través de tareas de desempeño. Un grado de dominio que no se ubica en la parte más alta de la progresión, por ejemplo, se representa mediante la identificación o selección de tareas o preguntas que solo están relacionadas de manera tangencial con el contenido, y por tanto, se dificulta la identificación de la relación entre estas y la habilidad que se quiere evaluar.

Posibles indicadores de evaluación o manifestaciones de esta dimensión para dar cuenta de la progresión en este ámbito, ordenados desde menor a mayor dificultad o complejidad de la tarea a desempeñar son los siguientes:

- Seleccionar preguntas, tareas, o rúbricas relacionadas con los contenidos evaluados, aunque pueden presentar falta de alineamiento con las habilidades.
- Identificar inconsistencias evidentes (problemas de validez) en la relación entre rúbricas, instrucciones, y objetivos de evaluación. Ej. la pregunta no se relaciona con el contenido a ser evaluado.
- Identificar inconsistencias sutiles (problemas de validez) en la relación entre rúbricas, instrucciones, y objetivos de evaluación). Ej. la pregunta está relacionada con el contenido, pero no con la habilidad.
- Seleccionar preguntas que son apropiadas para evaluar una determinada habilidad (o nivel de desafío cognitivo).

(b) ***La suficiencia de información que las tareas o estímulos seleccionados o planteados por el docente permiten generar***. Esto resulta crítico, tanto para la validez de las inferencias que se hagan a partir de los resultados o notas (por el grado en que el dominio ha sido cubierto) como para la confiabilidad de ellos. Así también, en otras circunstancias, la suficiencia de información apunta más a cubrir adecuadamente al grupo de estudiantes, para tomar decisiones pedagógicas atingentes para el grupo y no solo para unos pocos estudiantes que participan más activamente. Aquí nos interesará poder identificar cuán activo es el docente en recoger información sobre el aprendizaje, cuánta evidencia recoge, de cuántos estudiantes y con cuánta frecuencia.

Posibles indicadores de evaluación o manifestaciones de esta dimensión para dar cuenta de la progresión en este ámbito, ordenados desde menor a mayor dificultad o complejidad de la tarea a desempeñar son los siguientes:

- Diseñar instrumentos relacionados con el o los objetivos de aprendizaje.
- Formular preguntas a estudiantes específicos, deteniéndose cuando obtiene la respuesta correcta.
- Formular preguntas al curso en su conjunto para que respondan a coro.
- Diseñar instrumentos que ofrecen suficientes oportunidades para demostrar el aprendizaje de cada objetivo.
- Formular preguntas asignando aleatoriamente quiénes deben responder, o procurando que todos participen, hasta obtener una variedad de respuestas.

(C) El grado en que las preguntas e instrucciones planteadas por el docente están bien formuladas, suficientemente claras, directas, precisas, y libres de sesgos para favorecer que las respuestas o desempeño de los estudiantes efectivamente proporcionen la evidencia que se busca. Preguntas ambiguas, instrucciones confusas, uso de vocabulario que introduce fuentes de dificultad ajenas a lo que se busca evaluar, son algunos ejemplos de problemas frecuentes del modo en que el proceso de levantar evidencia sobre el aprendizaje puede verse interferido tanto en la interacción en aula como en el desarrollo de instrumentos formales. La hipótesis aquí es que en niveles incipientes de dominio los docentes podrán elaborar o seleccionar tareas cuya formulación se atenga a criterios muy básicos de formulación (por ejemplo, de claridad de la redacción o de las instrucciones) o de identificar errores de formulación muy evidentes. En cambio, en niveles avanzados hipotetizamos que los docentes podrán evaluar críticamente problemas de formulación menos evidentes o que se manifiestan a través de las respuestas de los estudiantes.

Posibles indicadores de evaluación o manifestaciones de esta dimensión para dar cuenta de la progresión en este ámbito, ordenados desde menor a mayor dificultad o complejidad de la tarea a desempeñar son los siguientes:

- Elaborar/seleccionar tareas cuya formulación se atiene a criterios básicos de formulación (ej. claridad, ausencia de errores conceptuales evidentes).
- Descartar tareas que presentan errores evidentes de formulación, al elaborar un instrumento, cuando selecciona las que utilizará.
- Elaborar tareas/ítems de evaluación cuya formulación es precisa y responde a condiciones más específicas (i.e. criterios de construcción de rúbricas, extensión de las opciones en preguntas de múltiple opción, etc.).
- Descartar o reformular tareas de evaluación que presentan errores menos evidentes de formulación, cuando selecciona las que utilizará.
- Evaluar críticamente tareas de evaluación (ej. su claridad) ante las respuestas de sus alumnos.

(d) ***El grado en que las tareas de evaluación que plantea el docente permiten identificar diferentes niveles de comprensión o de logro de un aprendizaje*** **y diagnosticar preconcepciones relevantes**. En el caso de instrumentos formales esto se manifestará en la habilidad del docente para seleccionar tareas o preguntas que permitan cubrir un amplio rango del aprendizaje evaluado, identificando las que se relacionan con diferentes niveles de logro, o bien en su capacidad para formular o seleccionar rúbricas que permitan describir tales niveles. En la evaluación formativa en sala de clases, esto se manifestará en tareas que también permitan evidenciar esta diversidad y no se limiten a preguntas retóricas o que solo apuntan a distinguir entre acierto y error. Esta habilidad es crítica pues permitirá a los docentes diseñar estrategias pedagógicas que respondan a la diversidad presente en su curso. Incluimos aquí también el grado en que las tareas evaluativas planteadas por el docente le permiten diagnosticar preconcepciones relevantes para el aprendizaje que busca desarrollar, y favorecen que los estudiantes hagan explícito su razonamiento (Heritage, 2011).

Posibles indicadores de evaluación o manifestaciones de la dimensión para dar cuenta de la progresión en este ámbito, ordenados desde menor a mayor dificultad o complejidad de la tarea a desempeñar son los siguientes:

- Identificar o proponer preguntas o ejercicios que resultan ser limitados en su potencial para recoger información, (por ser cerrados, retóricos, de

completación, o bien por indagar sobre conocimientos previos que se relacionan muy indirectamente con el aprendizaje a desarrollar).
- Crear o seleccionar instrumentos, rúbricas o conjuntos de preguntas que permiten describir distintos niveles de desempeño.
- Identificar o proponer preguntas que indagan y permiten revelar el razonamiento de los estudiantes, o acceder a información sobre preconcepciones relevantes para el aprendizaje a desarrollar.

2. Los criterios que hipotetizamos permitirían diferenciar distintos niveles de la calidad del desempeño de un docente en el ámbito de **analizar e interpretar evidencias** de aprendizaje:

(a) El grado en que las interpretaciones que el docente hace de las producciones y respuestas de un estudiante reflejan que el docente puede empatizar cognitivamente con él o ella y que presenta una comprensión (o al menos una hipótesis) sobre el razonamiento del estudiante (o se limitan a identificarlas como erróneas).

Posibles indicadores de evaluación o manifestaciones de esta dimensión para dar cuenta de la progresión en este ámbito, ordenados desde menor a mayor dificultad o complejidad de la tarea a desempeñar son los siguientes:

- Explicar las respuestas inesperadas o erradas de los estudiantes recurriendo elementos conductuales del estudiante (como no poner suficiente atención a las instrucciones).
- Formular explicaciones superficiales sobre el razonamiento que llevó al estudiante a una respuesta inesperada, errónea o incompleta.
- Formular explicaciones profundas sobre el razonamiento que llevó al estudiante a una respuesta inesperada, errónea o incompleta.

(b) El grado en que el análisis que hace el docente de las respuestas o producciones de un estudiante, o grupo de estudiantes, le permite ubicarlo acertadamente en un continuo de aprendizaje o nivel de profundidad de la comprensión de aspectos nucleares de la asignatura.

Posibles indicadores de evaluación o manifestaciones de esta dimensión para dar cuenta de la progresión en este ámbito, ordenados desde menor a mayor dificultad o complejidad de la tarea a desempeñar son los siguientes:

- Ubicar las producciones de los estudiantes en niveles de desempeño, sin relacionar el ordenamiento con una progresión del aprendizaje en desarrollo.
- Ubicar las producciones de los estudiantes en niveles de desempeño con algunas imprecisiones en este ordenamiento que reflejan una incipiente claridad de cómo progresa ese aprendizaje.
- Ordenar acertadamente las producciones de sus estudiantes en un continuo de aprendizaje y explicar este ordenamiento sobre la base de una progresión.

(c) **El grado en que la interpretación que hace el docente de información o datos producto de evaluaciones de aula de sus estudiantes** (informes de notas de un curso o grupo de cursos, informes de evaluaciones externas, entre otros), le permite formular hipótesis sobre los resultados de sus estudiantes.

Posibles indicadores de evaluación o manifestaciones de esta dimensión para dar cuenta de la progresión en este ámbito, ordenados desde menor a mayor dificultad o complejidad de la tarea a desempeñar son los siguientes:

- Interpretar información explícita en tablas o gráficos que dan cuenta de resultados de mediciones del aprendizaje.
- Interpretar información implícita en tablas o gráficos que dan cuenta de resultados de mediciones del aprendizaje.
- Formular hipótesis sobre los resultados de sus estudiantes sustentadas en características individuales o familiares de sus estudiantes.
- Formular hipótesis sobre los resultados de sus estudiantes analizando en forma coherente la enseñanza y toma decisiones.

3. Los criterios de progresión hipotetizados que permitirían diferenciar distintos niveles de la calidad del desempeño de un docente en el ámbito de **retroalimentar formativamente son**:

(a) El grado en que las intervenciones del docente son evaluativas (enjuiciadoras) o descriptivas, desde respuestas de aprobación/desaprobación, hasta entregar información descriptiva o preguntas que promuevan la indagación del estudiante sobre su desempeño.

Posibles indicadores de evaluación o manifestaciones de esta dimensión para dar cuenta de la progresión en este ámbito, ordenados desde menor a mayor dificultad o complejidad de la tarea a desempeñar son los siguientes:

- Reprobar las respuestas incorrectas y aprobar las incorrectas.
- Proporcionar la respuesta correcta ante respuestas erradas.
- Indicar dónde encontrar la respuesta correcta ante respuestas erradas.
- Entregar información nueva que permite ampliar la compresión del estudiante ante respuestas correctas o parcialmente correctas.
- Describir los logros que el estudiante ha alcanzado y sus errores e indicar formas de profundizar o mejorar.
- Promover la metacognición del estudiante para que sea él quien identifique sus logros y desaciertos y encuentre formas de profundizar o mejorar,
- Pedir al estudiante que utilice criterios de evaluación explícitos para monitorear y modificar su desempeño.

Es importante hacer notar que el potencial formativo de una retroalimentación no necesariamente es función de la medida en que esta es enjuiciadora o descriptiva, pues ello dependerá de la tarea en cuestión y del estudiante o el momento del proceso del aprendizaje. Por ejemplo, si un estudiante está aprendiendo la técnica para encestar más efectivamente en básquetbol, puede ser enteramente inadecuado que la retroalimentación del docente realice una retroalimentación en que pida al estudiante hacer una autoevaluación o descubrir qué está haciendo bien o mal. Por ello, el diseño de las tareas y de las opciones debe cuidar esta consideración.

(b) El grado en que el docente presenta modelos del desempeño esperado, como forma de retroalimentación, compartiendo criterios de evaluación, guiando el proceso de juicio y de definición de formas de profundización o mejora.

Posibles indicadores de evaluación o manifestaciones de esta dimensión para dar cuenta de la progresión en este ámbito, ordenados desde menor a mayor dificultad o complejidad de la tarea a desempeñar son los siguientes:

- Brindar la oportunidad para que los estudiantes discutan sobre el logro de la tarea y reflexionen sobre formas de mejorar, aunque los criterios de evaluación no sean explícitos.
- Entregar un modelo de desempeño esperado, el cual es contrastado por los estudiantes con su producto o tarea, aunque los criterios de evaluación no sean explícitos.
- Entregar un modelo de desempeño esperado y explicar por qué está logrado sobre la base de criterios de evaluación explícitos para que los estudiantes contrasten con su producto o tarea.
- Entregar modelos de distinto nivel de desempeño y favorecer la discusión sobre qué está logrado y no en cada uno, sobre la base de criterios de evaluación explícitos para que los estudiantes juzguen sus productos según estos criterios.

4. Proponemos que el desempeño del docente puede describirse como más o menos logrado según los siguientes criterios de progresión, en el ámbito de **certificar o calificar el aprendizaje alcanzado**:

(a) **El grado de validez y confiabilidad del juicio sobre el aprendizaje alcanzado:** en qué medida el juicio global realizado da cuenta del aprendizaje evaluado y ofrece un marco unívoco de interpretación. Aquí consideraremos, por una parte, la cobertura y relevancia de los aspectos que se consideran para juzgar el desempeño, las tareas que se seleccionan para evidenciarlo y la coherencia con su referente (objetivos de aprendizaje).

Posibles indicadores de evaluación o manifestaciones de esta dimensión para dar cuenta de la progresión en este ámbito, ordenados desde menor a mayor dificultad o complejidad de la tarea a desempeñar son los siguientes:

- Calificar con evidencia insuficiente para dar cuenta del aprendizaje evaluado o asignando igual peso a criterios independientemente de su relevancia para este.
- Certificar utilizando criterios que pondera según su relevancia para interpretar el nivel de logro alcanzado.
- Hacer un juicio global sobre el logro alcanzado, considerando evidencia relevante y suficiente para dar cuenta del aprendizaje evaluado (ej.: para llegar a la calificación de una unidad o año se consideran varias tareas consistentes con los aspectos esenciales del aprendizaje evaluado).

(b) ***El grado de comprensión de las reglas que subyacen a la asignación de notas.*** Como hipótesis, consideraremos más avanzados a los docentes que pueden dar cuenta de estas reglas y comprenden los principios que subyacen a los clásicos algoritmos utilizados por softwares y reglamentos de evaluación (por ejemplo, el 60% de logro, el uso de varias fuentes de evidencia o la necesidad de basar una calificación final en varias instancias de evaluación, etc.) y, por lo tanto, pueden cuestionar prácticas y tomar decisiones criteriosas con mayor flexibilidad.

Posibles indicadores de evaluación o manifestaciones de esta dimensión para dar cuenta de la progresión en este ámbito, ordenados desde menor a mayor dificultad o complejidad de la tarea a desempeñar son los siguientes:

- Usar softwares o aplicaciones para calcular notas conociendo las reglas de cálculo de ellas, aunque su comprensión de los supuestos que les subyacen puede ser limitada.
- Evaluar críticamente prácticas habituales de calificación, interpreta y cuestiona el significado de las notas en cuanto al aprendizaje de los estudiantes (ej. la pertinencia de promediar notas para obtener una de síntesis, o lo relativo de asignar el 4 al 60% de logro).

(c) **El Grado de transparencia de los criterios en las prácticas de calificación:** grado en que la calificación ha sido previamente planificada y se vela porque los estudiantes tengan claridad de los criterios de calificación (rúbricas, pautas), ya sea antes o después de asignar la nota.

- Entregar información procedimental sobre las calificaciones (ej. su planificación en el calendario, la ponderación de cada una, el tipo de instrumento que se usará).
- Calificar según criterios claros preestablecidos y conocidos por los estudiantes (ej. los estudiantes conocen la rúbrica con que se calificará su trabajo y el peso que tendrá cada criterio).

Características de los instrumentos

Cuestionario Sociodemográfico: se incluyó un grupo de preguntas como sexo, año de nacimiento, formación inicial, años de experiencia, especialización en evaluación, dependencia administrativa, región del establecimiento donde se desempeña la mayor parte del tiempo, y asignatura(s) que enseña, entre otras variables que permitieron tener una caracterización más profunda de los participantes.

A partir de la revisión bibliográfica (ej. Gotch & French, 2014), se determinó que los principales instrumentos utilizados para medir habilidades esta área, son cuestionarios de autorreporte y pruebas de selección múltiple en las que se evalúa los conocimientos de los docentes sobre evaluación.

Este estudio contempla la elaboración de dos instrumentos:

- El primero es un conjunto de tareas que apuntan a medir competencias en evaluación con ítems de selección múltiple y preguntas abiertas.
- El segundo es un cuestionario de autorreporte que busca recoger prácticas de evaluación de aprendizaje que realizan los docentes en el aula.

Construcción de los instrumentos piloto

En esta parte se abordan las acciones realizadas para la construcción de los instrumentos pilotos: construcción de los ítems; ensamblaje de las formas piloto y diseño de la plataforma online.

Construcción de los ítems o tareas para la prueba

Como resultado de las definiciones anteriores, la matriz para la construcción de la prueba de competencias evaluativas fue la siguiente:

Tabla 2. Matriz de construcción prueba piloto de competencias evaluativas

	DIMENSIONES			
DIFICULTAD (estimada)	Recoger evidencia del aprendizaje	Analizar e interpretar evidencia del aprendizaje	Retroalimentar formativamente	Certificar o calificar el aprendizaje alcanzado
Baja	4	4	4	4
Media	4	4	4	4
Alta	4	4	4	4
TOTAL	12	12	12	12

Describiremos a continuación el proceso de elaboración de la sección de las tareas e ítems para diagnosticar el desempeño de los docentes a través de sus respuestas y producciones ante los casos presentados.

a) Elaboración de casos, ítems y tareas

Los ítems combinaron preguntas cerradas, especialmente de selección múltiple y preguntas abiertas o de respuesta construida ante los casos presentados bajo la forma de descripciones de situaciones de aula, videos breves, instrumentos, citas de intervenciones de estudiantes o de interacciones profesor-alumno, etc.

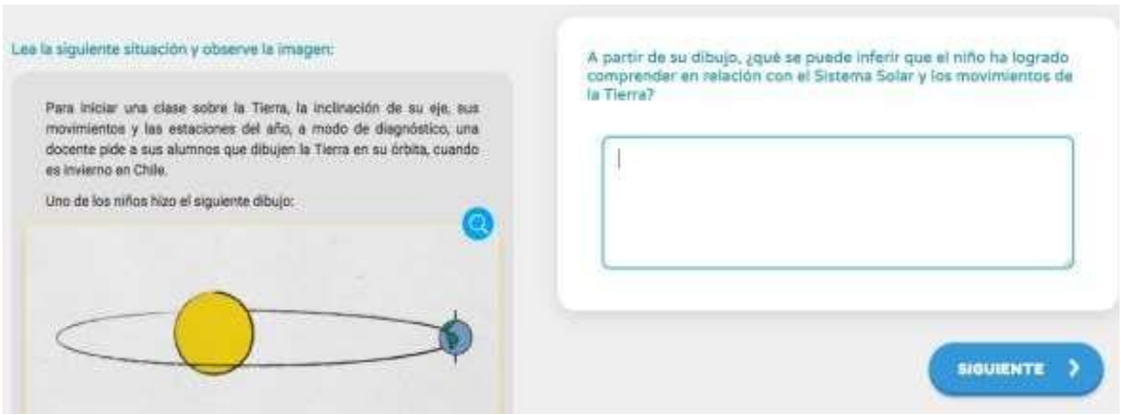
Los ítems fueron elaborados por el equipo de investigadoras, siendo sometidos a numerosas iteraciones entre sus miembros. Posteriormente, se realizó una consulta a expertos en evaluación de aula y especialistas en didácticas de la enseñanza de las asignaturas de Matemática, Lenguaje, Ciencias e Historia, Geografía y Ciencias Sociales. El perfil de los expertos en evaluación corresponde a especialistas en el área, quienes revisaron aspectos técnicos de construcción de las preguntas y sus rúbricas. Estos profesionales fueron seleccionados considerando su participación en procesos similares y su expertise en el área de evaluación. Los expertos en evaluación juzgaron la correspondencia de los ítems construidos con la dimensión que dicen evaluar, en tanto, los expertos en las didácticas de las disciplinas verificaron además que los ítems no tuvieran errores conceptuales y qué sus contextos correspondieran con situaciones auténticas, es decir, situaciones con las cuales los docentes se encuentran en sala de clases.

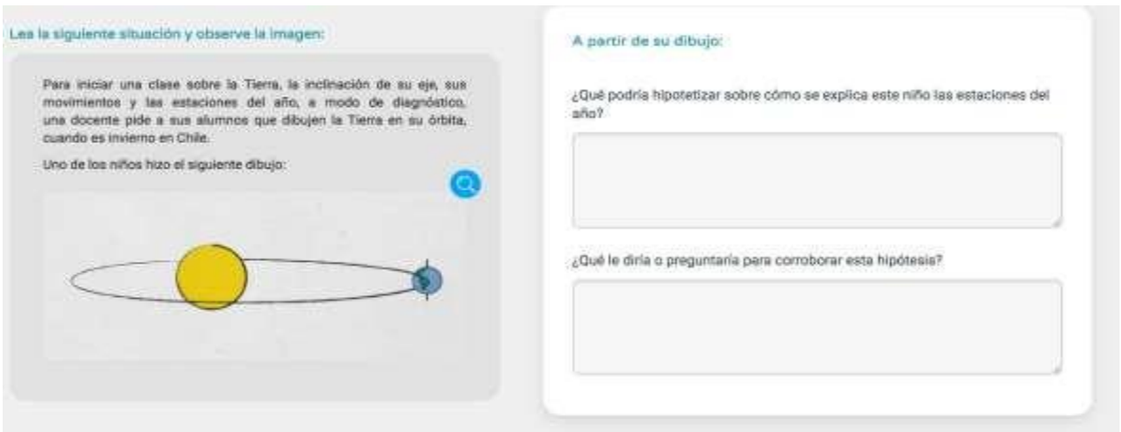
Además de los juicios de expertos en evaluación y didáctica, se realizaron entrevistas cognitivas (Howell, Phelps, Croft, Kirui & Gitomer, 2013; Young, King, Cogan, Ginsburgh, Kotloff, Cabrera, Cavalie, 2014) a una muestra de 7 estudiantes de educación básica prontos a su egreso. Ellos fueron citados a una jornada de trabajo, en la cual cada participante respondió las preguntas y junto con eso registró en una planilla la dificultad de la tarea solicitada y toda observación, dificultad o comentario que surgiera al momento de realizar la actividad. Una vez concluido esto, cada participante se reunió con un miembro del equipo de investigación y procedió a relatar sus impresiones, los posibles errores o dificultades detectados, como asimismo, para dar cuenta del razonamiento que los había llevado a marcar tal o cual opción y a descartar las otras.

Con todas las observaciones recogidas, provenientes tanto de los expertos como de los futuros docentes que participaron en la entrevista cognitiva, ser realizaron ajustes tanto a las preguntas como a sus rúbricas. De este modo, de los 54 ítems inicialmente construidos, 42 fueron aprobados. Estos 42 ítems se distribuyen de forma balanceada en las 4 dimensiones definidas que componen la competencia evaluativa (ver en anexo la distribución de las tareas que se están piloteando en las dimensiones, en la 'Planilla maestra').

Las preguntas abiertas o de respuesta construida fueron 4, una por cada dimensión evaluada. Estas buscaban evaluar habilidades tales como crear una retroalimentación, producir una interpretación de la respuesta de un estudiante, o argumentar una decisión de calificación.

A continuación, se ilustra a través de un ejemplo de ítem de respuesta abierta, el tipo de tarea que se presenta en el instrumento. Las tres preguntas formuladas están referidas a la capacidad del docente para interpretar o analizar evidencia del aprendizaje.





El conjunto de tareas piloto en su versión online ya diagramada puede revisarse en su versión en Word en el anexo 'Pilotaje' específicamente en la carpeta 'Ítems'.

b) Elaboración del cuestionario de autorreporte de prácticas de evaluación

En forma complementaria a las tareas antes descritas, se elaboró un cuestionario de autorreporte de prácticas de evaluación de aula, para indagar sobre aspectos de las dimensiones menos evaluables a través de un instrumento que simula situaciones de sala de clases, o que hubieran requerido un desarrollo excesivamente sofisticado para ello. Por ejemplo,

interesaba conocer si al recoger evidencia, el docente busca información de la mayor parte de los estudiantes o si por el contrario detiene su búsqueda cuando obtiene la respuesta correcta. Reproducir esta situación en una tarea de desempeño online hubiera resultado muy artificial o bien extremadamente complejo pues hubiera requerido un diseño interactivo. Considerando lo anterior, se complementó las tareas e ítems con un cuestionario de autorreporte en que se consulta al profesor por sus prácticas más habituales.

El cuestionario consistió en una serie de afirmaciones frente a las cuales el docente debía marcar la frecuencia con que realiza las prácticas descritas (ver el cuestionario en Anexo carpeta 'Piloto' subcarpeta 'cuestionario autorreportado).

La tabla de especificaciones del instrumento piloteado fue la siguiente:

Cuestionario	Recoger evidencia	Analizar e interpretar	Retroalimentar	Certificar
Nº ítems	10	2	12	13

A continuación algunos ejemplos de reactivos de este cuestionario (Nunca o casi nunca/Algunas veces/Frecuentemente/Siempre o Casi siempre).

- Sorteo o asigno al azar quiénes deben responder las preguntas que. formulo.
- Doy la palabra solo a quienes se ofrecen a responder o levantan la mano.
- Realizo preguntas a mis estudiantes hasta que alguno responde correctamente.
- Uso estrategias para conocer la respuesta de todos o casi todos mis estudiantes (ej. que escriban su respuesta en un papel que recojo, o que 'voten' por una respuesta).
- Termino la clase haciendo una pregunta o ejercicio breve que todos deben responder por escrito y entregarme.
- Pido a los estudiantes que respondieron lo que yo esperaba, que expliquen cómo llegaron a su respuesta
- Hago preguntas generales para indagar sobre conocimientos previos (Ej. ¿quién me puede decir qué trabajamos la clase anterior?)

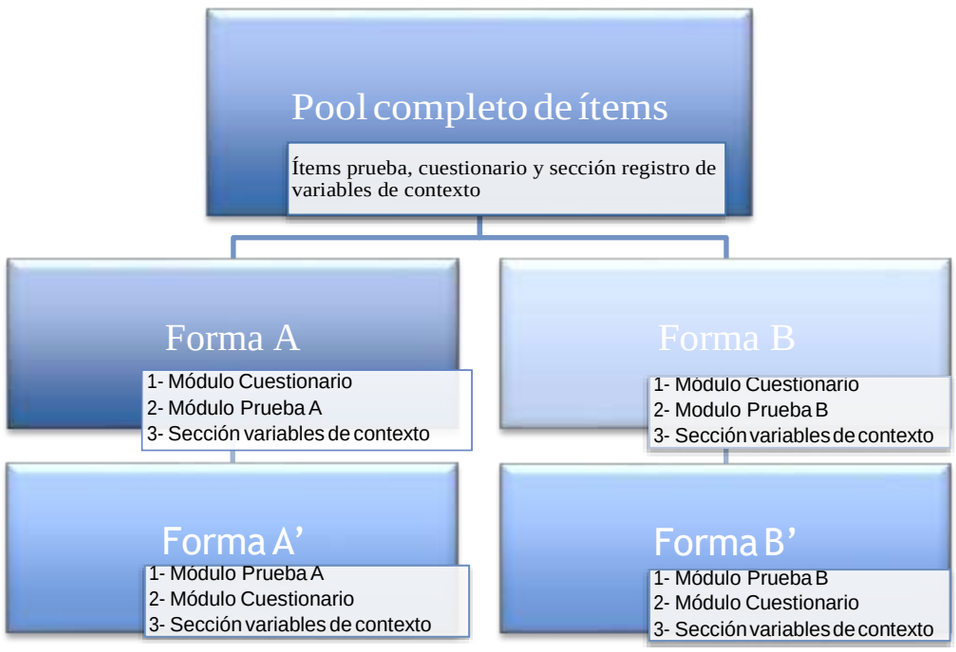
Ensamblaje del instrumento piloto y Diseño plataforma web

Después de la elaboración de los ítems y del cuestionario, se diagramaron las tareas, se montó el instrumento en una plataforma virtual, y se probó su funcionalidad. También se implementaron las medidas para orientar las respuestas de los docentes con menos habilidades TIC, incluyendo controles de retroalimentación del avance, y un mecanismo para que los profesores que no disponen del tiempo para completar el instrumento en una oportunidad puedan completarlo en un momento posterior o inclusive en varias sesiones (así, se espera obtener una mayor tasa de respuesta).

La elaboración de la plataforma web para la aplicación online del instrumento y la posterior generación de reportes automáticos estuvieron a cargo del Instituto de Informática Educativa de la Universidad de la Frontera de Temuco.

Finalmente, en un trabajo conjunto entre el equipo UC, Universidad de los Andes y UFRO; se distribuyeron y ensamblaron las 42 tareas en cuatro formas piloto de 24 preguntas (A y B y sus respectivos espejos A' y B', con una distinta ordenación de los reactivos[2]) que se asignaron aleatoriamente a quienes ingresaron a la plataforma para responder. Por lo tanto, el instrumento completo fue piloteado en 4 formas (ver organización interna de las formas piloto en Figura 2). Las formas A y B se ensamblaron resguardando que fueran paralelas, (equilibrándolas en cuanto al número de preguntas, su contenido, las dimensiones a las que responden y su carga lectora).

Figura 2. Esquema de las formas de los instrumentos



La plataforma tecnológica utilizada para la aplicación piloto de Mejor Evaluación corresponde a un sistema Web basado en *LimeSurvey*[3], una aplicación de código abierto que permite crear, editar y adaptar instrumentos de evaluación, así como gestionar aplicaciones masivas vía Internet y almacenar los datos asociados a las respuestas de los participantes.

La plataforma utilizó una arquitectura tecnológica de tres capas: a) ***capa usuario***, que es el ambiente de acceso de los profesores que participaron de la evaluación, para lo cual se usa un navegador Web instalado en el computador de cada usuario (Chrome, Firefox o similar), b) ***capa de aplicaciones***, consistente en el servidor Web donde se encuentran instalados el módulo de autentificación, módulo de administración de instrumentos y los módulos que apoyan la aplicación masiva, y c) ***capa de datos***, consistente en la base de datos en la cual se almacenan los registros de cada uno de los profesores participantes de la evaluación (datos demográficos, respuestas, tiempos de evaluación, etc.).

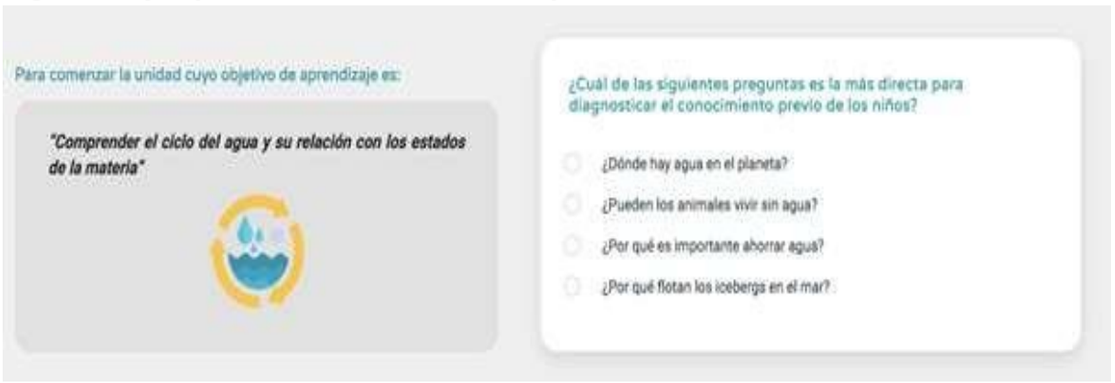
[2] Las formas A' y B' son idénticas a las formas A y B en cuanto a preguntas que contienen, pero se alterna la presentación del cuestionario o escala de autorreporte de prácticas evaluativas: En las formas A y B el cuestionario está situado a continuación de la prueba y en las formas A' y B' este se encuentra al principio de la misma.
[3] Ver Limesurvey en https://www.limesurvey.org/

La plataforma de evaluación piloto quedó disponible en www.mejorevaluacion.cl (ambiente web), al cual los profesores podían ingresar usando su RUN. (ver Figura 3 a continuación):

Figura 3. Acceso a plataforma piloto de evaluación

Una vez que los profesores ingresaban a la plataforma se les presentaba una colección de ítems relacionados con procesos de evaluación y retroalimentación en aula, los cuales contemplaban diversos tipos de estímulos (textos, imágenes y animaciones) frente a los cuales los profesores debían seleccionar una de las respuestas disponibles (ver figura 4), jerarquizar o bien escribir una respuesta abierta breve. A medida que los profesores iban respondiendo a las preguntas podían avanzar a las preguntas siguientes. Cabe señalar que los profesores tenían la opción de salir de la evaluación y volver a entrar con su RUN para continuar y finalizar la evaluación desde el último ítem que habían respondido.

Figura 4. Ejemplo de ítem selección múltiple



Junto con la prueba, la plataforma de evaluación digital incluye un consentimiento de participación, un cuestionario de auto reporte y un cuestionario sociodemográfico que deben ser respondido directamente en la plataforma. La modalidad de aplicación fue de tipo libre e independiente, esto implica que los profesores fueron invitados virtualmente a participar de la evaluación enviándoles el link de la plataforma, donde podían acceder y responder en forma autónoma.

A medida que los profesores ingresaban e iban respondiendo a las preguntas de la evaluación, sus respuestas iban siendo almacenadas en un servidor centralizado en Internet, que les permitía tener acceso a las planillas de respuestas para realizar los análisis.

También se desarrolló una versión de la plataforma que se pudiera responder en Tablet y en celular, de manera que fuera posible hacer un operativo de campo presencial utilizando alguno de estos equipos.

En forma paralela a la plataforma de aplicación de la evaluación se desarrolló un sistema de monitoreo de la implementación, que permitía ir revisando en tiempo real la participación de los profesores y la distribución de las principales variables de agrupación (ver Figura 5 a continuación).

Figura 5. Sistema de Monitoreo de la Aplicación

6.2. Etapa 2: Aplicación piloto de los instrumentos

Muestra de docentes para la prueba piloto

La población de estudio la constituyen todos los profesores de educación básica de primero a sexto año de educación básica y que se encuentran actualmente ejerciendo la profesión en las escuelas de Chile. Las fuentes de información corresponden a los docentes que conforman la muestra. Se trata de la aplicación de un instrumento, cuya duración es de aproximadamente una hora.

La muestra tuvo el propósito de contar con variabilidad en términos de conocimientos y habilidades en el ámbito de la evaluación de aprendizajes, para poder probar el instrumento y poder describir niveles crecientes de desempeño a partir de sus resultados. No era el propósito de este muestreo el representar la población de profesores para obtener resultados que puedan generalizarse a la población, sino disponer de la máxima variabilidad posible en cuanto a representación de zonas geográficas de Chile y de dependencia de los establecimientos donde los docentes se desempeñan la mayor parte del tiempo.

Los participantes del estudio se seleccionaron mediante un muestreo no probabilístico intencionado, pues su propósito fue contar con una muestra variada en conocimientos y habilidades en el ámbito de la evaluación de aprendizajes de los docentes; representación de zonas geográficas de Chile y tipo de administración de los establecimientos donde los profesores se desempeñan.

Ingresaron a la plataforma 1040 profesores, 852 respondieron de manera íntegra o solo una parte de ambos instrumentos. Se eliminaron de la muestra aquellos docentes que contestaron menos de la mitad de las preguntas, quedando 545 que respondieron las tareas de evaluación y 608 el cuestionario. Teniendo en cuenta la extensión de ambos instrumentos, también se eliminaron de la muestra aquellos docentes que demoraron menos de 30 minutos en responder. Finalmente, y con el propósito de contar con patrones completos de respuestas para realizar escalamiento IRT, la muestra definitiva estuvo conformada por 398 docentes de educación primaria que respondieron la totalidad de las tareas en la plataforma online.

De los 398 profesores 193 respondieron la forma A y 205 la forma B. En cuanto a sexo, 85% declaran ser mujeres (n=339) y 15% hombres (n=59). Respecto de la zona geográfica, el 50% reporta ser de la Región Metropolitana (n=197) y 50% del resto de las regiones de Chile (n=201). El 38% indica tener de 1 a 5 años de ejercicio docente (n=150), 28% entre 6 y 11 años (n=113), 25% entre 11 y 25 años (n=98) y 9% 25 años o más (n=37). En relación con el tipo de establecimiento educacional donde se desempeñan los profesores, 36% reporta hacerlo en el sector público (n=143), 45% en el particular subvencionado (n=178) y el 19% en el particular pagado (n=77).

Aplicación del Instrumento piloto

Convocatoria abierta

La aplicación del instrumento se realizó principalmente a partir de una convocatoria abierta en redes sociales y a través de bases de profesores (Educar Chile, Red de profesores de Chile, Facebook, bases de datos de Alumni de pedagogía básica de distintas universidades, etc.) donde se publicó la invitación mediante un banner para responder la prueba online (ver figura 6), o bien se envió un correo masivo también con un link directo para entrar a la plataforma (ver figura 7). Como una forma de incentivar la respuesta del instrumento en su totalidad, se realizó un sorteo de premios (giftcard de una librería o casa comercial).

Figura 6. Banner clikeable para responder instrumento online

Figura 7. Correo masivo con link directo

Los profesores que ingresaron a la plataforma respondieron al comienzo algunas preguntas de contexto, tales como la región y la dependencia del establecimiento en que se desempeñaban. Con esta información la plataforma dirige al profesor a una de las formas, de manera de equilibrar las respuestas de los docentes y no tener una región o una dependencia sobrerrepresentada.

Asimismo, los docentes entregaron su correo electrónico, de este modo se pudo enviar cada 15 días recordatorios a aquellos profesores que ingresaron a la plataforma pero que no concluyeron la tarea. Cuando el docente reingresaba, la plataforma lo dirigía al sitio exacto hasta donde había alcanzado a contestar.

Como medida de contingencia, para alcanzar un tamaño muestral mínimo de 400 casos y Para reducir el sesgo de autoselección que pudiera ocasionar la estrategia de convocatoria abierta, se realizó un operativo de campo en la sede de Villarrica de la Facultad de Educación de la Pontificia Universidad Católica.

Operación de campo complementaria

Para incrementar el número de participantes y aumentar la variabilidad de la muestra de profesores en el piloto, se realizó un operativo de campo con un grupo de profesores de escuelas de la zona sur del país. Estos profesores asistieron a un curso en la Facultad de Educación del Campus Villarrica de la Universidad Católica y respondieron libremente el instrumento.

Los profesores accedieron a la plataforma online a través de Tablet previamente configuradas y conectadas a internet, y contaron con aproximadamente 90 minutos para responder la totalidad del instrumento.

6.3. Etapa 3: Análisis de datos y ensamblaje de instrumentos definitivos

Análisis de datos psicométricos de la aplicación piloto

En este apartado se detallan los análisis psicométricos llevados a cabo para cada uno de los instrumentos aplicados. En el cuestionario de autorreporte de prácticas evaluativas los análisis se enfocaron en la depuración del mismo a fin de construir una versión final. En el caso de la prueba de competencias, los análisis abarcaron dos aspectos diferentes: detectar el conjunto de ítems máximamente confiable y válido con el fin de reportar resultados a quienes participaron de la aplicación piloto y, por otra parte, seleccionar un conjunto de ítems con buenas características psicométricas para constituir un instrumento definitivo que posteriormente quedará disponible para ser utilizado.

Prueba de competencias evaluativas

Se realizaron análisis clásico de ítems, incluyendo distribución de respuestas, grado de dificultad[4], discriminación[5] y se construyeron las curvas empíricas. Para el cálculo de la confiabilidad se utilizó el indicador de alfa de Cronbach.

Se realizó un análisis factorial exploratorio por componentes principales[6], método canónico para la realización de análisis exploratorios, con el fin de determinar la unidimensionalidad de la prueba y poder estimar un resultado. Se utilizó la rotación varimax, método más utilizado cuando se trata de rotaciones ortogonales (los factores obtenidos no correlacionan entre ellos) y de fácil interpretación.

Para la puntuación de los sujetos y la construcción del instrumento definitivo se llevó a cabo un análisis IRT bajo el modelo de Rasch, que considera solo el parámetro de dificultad para la caracterización de cada ítem. Este modelo permite generar una equivalencia entre la cantidad de respuestas correctas y el puntaje reportado. El método Rasch resulta ampliamente utilizado en el análisis de instrumentos porque requiere una menor cantidad de respuestas para llevar a cabo el análisis según la Teoría de Respuesta al Ítem (IRT) y a la facilidad de interpretación de los resultados obtenidos. Este análisis se llevó a cabo con el software Winsteps.

[4] La dificultad clásica corresponde al porcentaje de sujetos que selecciona la alternativa correcta.
[5] La discriminación clásica corresponde a una correlación punto biserial.
[6] Cabe destacar que si bien el Análisis por componentes principales (APC) no es un análisis factorial puro puesto que su propósito es reducir la dimensionalidad de las variables consideradas y no buscar un factor subyacente (como sucede en el Análisis Factorial), la ventaja de utilizar APC es la flexibilización del criterio de normalidad de las variables estudiadas.

Cuestionario de autorreporte de prácticas evaluativas

Se realizaron análisis de la frecuencia de respuesta y de correlación de los ítems del cuestionario de autorreporte. También se calculó el alfa de Cronbach y se llevó a cabo un análisis por componentes principales, para determinar la existencia de un factor preponderante para la construcción del instrumento definitivo.

Adicionalmente se realizó un análisis IRT con el modelo de respuesta graduada Andrich en el software Winsteps, para realizar un escalamiento de los ítems y analizar el perfil o patrón general de respuestas.

Fase de ensamblaje del instrumento definitivo

El procedimiento para seleccionar los ítems y ensamblar la prueba definitiva consideró los siguientes criterios: el grado de dificultad, la cobertura de las dimensiones evaluadas y la variedad de tipo de tarea o formato de ítem de evaluación, esto último solo en el caso de la prueba.

A partir del análisis de las respuestas de docentes en la aplicación piloto se ensambló el instrumento definitivo y se ordenaron las preguntas según su nivel de dificultad empírico (de menor a mayor desempeño). El propósito original de este ordenamiento era distinguir niveles de logro estableciendo un puntaje de corte que permitiera distinguir categorías de desempeño. Sin embargo, dados los índices de confiabilidad del instrumento, esto no fue posible debido al error de clasificación asociado al error de medición de los puntajes. Sin embargo, se realizó una descripción cualitativa de las habilidades requeridas por las tareas contenidas en el instrumento para diferentes 'zonas' de la prueba. Estas descripciones se utilizarán en el reporte de resultados, como se muestra más adelante.

7. RESULTADOS

En este apartado se incluyen los principales resultados de los análisis psicométricos de los instrumentos, las decisiones adoptadas para reportar resultados de la aplicación piloto y conformar los instrumentos definitivos. Finalmente, se reportan los resultados en términos de productos obtenidos (los instrumentos online).

7.1. Prueba piloto de competencias evaluativas: selección de preguntas para el reporte piloto

En el caso de la prueba piloto, se disponía de un pool de 42 preguntas en total, 24 en la forma A (alfa de Cronbach = 0,53) y 24 en la forma B (alfa de Cronbach = 0,50), con 5 ítems comunes entre ellas: tres de selección múltiple, una de respuesta abierta y una de jerarquización. De las 42 preguntas, 8 fueron descartadas de los análisis psicométricos por problemas de planteamiento detectados a posteriori.

Para las 34 preguntas restantes se obtuvo la frecuencia de respuesta de cada opción y la correlación opción-test, la discriminación (a partir de la correlación punto biserial) y el alfa de Cronbach. En el primer análisis de estos ítems se obtiene un alfa de 0,60, con un número de respuestas correctas promedio de 10 y con una dificultad clásica promedio de 52%. La discriminación clásica promedio de las preguntas fue de 0,30.

Como resultado del análisis factorial exploratorio, con método de componentes principales y rotación Varimax, se extraen tres factores. Considerando que para trabajar con modelo IRT, Rasch en este caso, el requisito de base es que el instrumento mida una única dimensión subyacente o factor, por ello se decidió realizar una selección de los ítems a partir de los resultados métricos, principalmente el AFE (Tabla 3).

Tabla 3. pesos factoriales de los 24 ítems que cargan en un único factor

Dimensión	Forma	Rótulo	Posición	Peso factorial
Recoger evidencia del aprendizaje	A	Rec_C_cicloagua	p1	0,364
		Rec_M_lápices	p6	0,334
		Rec_L_tarta _1	p17	0,294
		Rec_M_manzanas	p24	0,308
		Rec_L_noticias	p32	0,466
Analizar e interpretar evidencia del aprendizaje		Ana_L_tarta_4	p18	0,366
Retroalimentación formativa		Retro_M_rectángulo	p7	0,607
		Retro_L_fábula	p9	0,381
		Retro_M_gráfico1	p26	0,476
		Retro_H_plaza Italia	p28	0,276
		Retro_C_mamut	p30	0,424
		Retro_C_circuito	p31	0,294
		Retro_C_fotosintesis	p_abierta	0,336
Certificar o calificar el aprendizaje		Cer_M_plumeros_1	p12	0,499
		Cer_libreta de notas	p15	0,334
Analizar e interpretar evidencia del aprendizaje	B	Ana_M_fracciones figuras	p2b	0,337
		Ana_M_patron vertical	p28b	0,458
Retroalimentación formativa		Retro_M_gráfico2	p23b	0,318
		Retro_M_resta	p31b	0,554
Certificar o calificar el aprendizaje		Cer_M_plumeros (b)	p8b	0,435
		Cer_L_tortuga y arana	p10b	0,470
		Cer_C_no progresion de notas	p13b	0,350
		Cer_C_ciencias de la vida	p22b	0,428
		Cer_H_conquista espanola	p30b	0,439

Para llevar a cabo estos análisis, se combinaron las respuestas de ambas formas utilizando los ítems comunes o anclas. A partir de estas preguntas en común se realizó un enlace entre ambas formas con 24 ítems (de los 34 analizados), los que permitieron que los resultados de los sujetos quedasen en la misma escala para posteriormente hacer el reporte.

Con estas 24 preguntas seleccionadas de ambas formas, se realizó un análisis Rasch, generando una puntuación individual en una escala de 0 a 50 puntos y un mapa de ítems para poder describir los

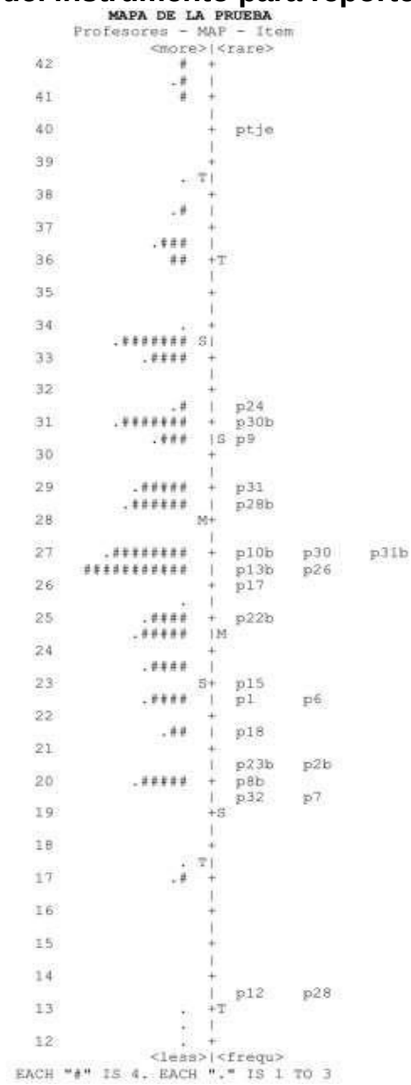
niveles alcanzados (ver figura 8), siendo la dificultad del instrumento de 25 puntos (24,77). Estas 24 preguntas reportan una confiabilidad de 0,64, con una dificultad promedio clásica de 60% y una discriminación promedio de 0,37.

Se decidió reportar resultados a quienes participaron del pilotaje del instrumento con este subconjunto de 24 preguntas, las que se distribuyen según dimensión y dificultad (tabla 4).

Tabla 4. Distribución de los 24 ítems seleccionados post pilotaje según dificultad IRT y dimensión evaluada

Dificultad obtenida	Recoger evidencia	Analizar e interpretar	Retroalimentar	Certificar
Baja	1	2	3	2
Media	3	0	3	4
Alta	1	1	3	1
Total	5	3	9	7

Figura 8. Mapa de ítems del instrumento para reporte de resultados del piloto



Adicionalmente, se utilizó el mapa de ítems de la Figura 8 para describir cualitativamente el desempeño en tres 'zonas' de la escala (alta, baja y mediana dificultad). Dado que la

confiabilidad del instrumento que se logró conformar no es suficiente, se optó por no describir niveles de logro ni establecer puntajes de corte entre categorías de desempeño. Asimismo, se estimó el error de medición a lo largo de la escala, para reportar, junto con el puntaje alcanzado, su intervalo de confianza, lo cual se expresó gráficamente en el reporte como una banda en torno al puntaje.

7.2. Diseño del reporte de resultados a los sujetos que respondieron versión piloto

Como se señaló, sobre la base de los 24 ítems seleccionados, se pudo estimar un puntaje para cada sujeto que respondió la prueba piloto. Para cada puntaje se estimó además el error de medición, por lo que el puntaje se reporta a cada sujeto con su intervalo de confianza, tal como se muestra en la Figura 9.

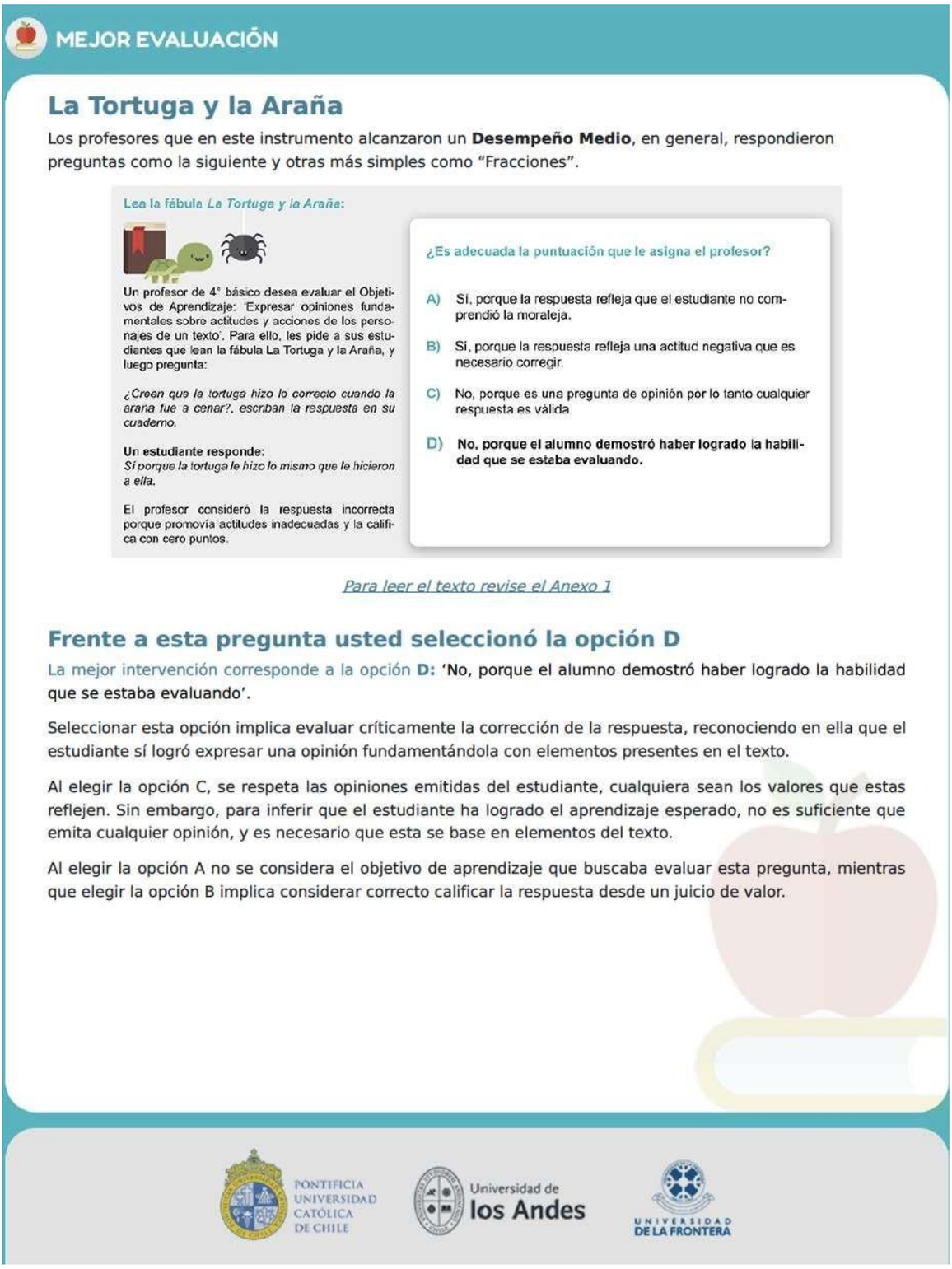
Figura 9. Ejemplo de intervalo de confianza de puntajes

Asociados a las 'zonas' de puntaje en la prueba, se desarrollaron descripciones basadas en los ítems ubicados en las respectivas zonas de dificultad de la prueba, por ejemplo:

> Quienes tuvieron un alto desempeño en este instrumento, en general respondieron adecuadamente tareas que requerían reaccionar a la respuesta de un estudiante reformulando las instrucciones entregadas para mejorar su precisión y claridad, discriminar intervenciones de retroalimentación según su potencial para promover la habilidad de inferir información de un texto, además de integrar y priorizar información para concluir sobre el aprendizaje alcanzado por los estudiantes. Además, en general, pudieron resolver adecuadamente las tareas descritas para el nivel 'medio' y 'bajo'.

El reporte incluye, además del puntaje y la descripción transcrita más arriba, resultados individualizados para 3 ítems, con una interpretación del significado de haber elegido cada una de las opciones, hayan sido correctas o incorrectas (ver figura 10).

Figura 10. ejemplo de resultado individualizado por ítem



Estos reportes fueron enviados, a partir del 19 de diciembre por correo electrónico a cada uno de los docentes que completaron el instrumento piloto.

7.3. Prueba definitiva online y reporte

Se continuó con el proceso de depuración del instrumento con el fin de construir su versión final, la que quedará disponible de manera electrónica para que pueda ser respondida por los docentes interesados en conocer una apreciación global de sus competencias evaluativas.

Dado el extenso tiempo de respuesta de la versión piloto (una hora para 24 preguntas), se decidió acortar el instrumento definitivo online a 20 ítems, escogiendo aquellos con los mejores parámetros psicométricos y que respondían de mejor modo al constructo evaluado. La distribución según dificultad IRT y dimensión evaluada de estas 20 preguntas se presenta en la Tabla 5.

Tabla 5. Distribución de ítems de prueba definitiva según dificultad IRT y dimensión evaluada

Dificultad obtenida	Recoger evidencia	Analizar e interpretar	Retroalimentar	Certificar
Baja	1	2	2	0
Media	3	1	2	3
Alta	1	1	3	1
Total	5	4	7	4

Los rangos de los parámetros psicométricos utilizados se presentan en la Tabla 6 (Costello & Osborne, 2005; Ebel, 1979; Hotiu, 2006).

Tabla 6. Rangos aceptables de parámetros psicométricos utilizados como criterios de selección de los ítems

	Dificultad clásica	Discriminación	Peso factorial
Mínimo	0,2	0,20 a 0,29	0,3
Máximo	0,9	No hay máximo	No hay máximo

Para efectos de cálculo de puntaje y generación de los reportes online, se utilizaron en total 13 preguntas de las 20 (solamente aquellas provenientes de la forma A). Esto, debido a que todas ellas fueron respondidas por los mismos sujetos en la aplicación piloto y por lo tanto, su escalamiento para la estimación de un puntaje resulta más confiable que si este se estimara en base a preguntas respondidas por distintos sujetos (en la forma A y B) y con un núcleo muy reducido de ítems ancla entre ambas formas. Los siete ítems restantes, que se incluyen en el instrumento online, provienen de la forma B, pero en un comienzo, no serán utilizados para el reporte online. Su inclusión tuvo por objeto robustecer la matriz de evaluación de la prueba, de modo que, a futuro, una vez que se cuente con datos de aplicaciones sucesivas del instrumento completo respondido por los mismos sujetos, será posible integrarlas a la estimación de los puntajes y luego a los reportes online.

Las 13 preguntas de la forma A utilizadas para el cálculo del puntaje definitivo presentan en conjunto una confiabilidad de 0,61, una dificultad IRT promedio de 24,88 y una discriminación punto biserial promedio de 0,33. Los docentes presentaron en promedio 5 respuestas correctas para este conjunto de ítems. Con el objetivo de facilitar la posterior lectura de los resultados, nuevamente la escala se fijó entre 0 y 50 puntos.

Del mismo modo que para los sujetos que respondieron el piloto, quienes completen la nueva versión de la prueba online, recibirán un reporte automático con su resultado. Las características de este

reporte serán muy similares a las del reporte antes presentado. El instrumento definitivo online se encuentra en el sitio web www.mejorevaluación.cl. La Tabla 7 muestra el flujo de decisiones en relación con la eliminación o retención de los ítems en las distintas etapas.

Tabla 7. Flujo de síntesis de decisiones de eliminación o retención de ítems

Rótulo	Progresión	Asignatura	Tipo	1er filtro: Problemas de planteamiento	2do filtro: Análisis factorial (AFE)	3er filtro: Extensión del instrumento
Ana_C_Traslación_1	Analizar	Ciencias	Abierta	X		
Ana_C_Traslación_2	Analizar	Ciencias	Abierta	X		
Ana_H_Continuidad y Cambio	Analizar	Historia	Opción múltiple		X	
Ana_L_Tarta_4	Analizar	Lenguaje	Opción múltiple			
Ana_L_Tarta_5	Analizar	Lenguaje	Jerarquización	X		
Ana_L_Tortuga y araña	Certificar	Lenguaje	Opción múltiple			
Ana_M_Fracciones Figuras	Analizar	Matemática	Opción múltiple			
Ana_M_Patrón Vertical	Analizar	Matemática	Opción múltiple			
Ana_M_Secemo el pan	Analizar	Lenguaje	Selección	X		
Cer_C_Caritas Felices	Certificar	Ciencias	Opción múltiple		X	
Cer_C_ciencias de la vida	Certificar	Ciencias	Opción múltiple			
Cer_C_No Progresión de Notas	Certificar	Ciencias	Opción múltiple			
Cer_H_civilizaciones	Certificar	Historia	Opción múltiple		X	
Cer_H_Conquista Española	Certificar	Historia	Opción múltiple			
Cer_H_Porcentaje de logro	Certificar	Historia	Opción múltiple		X	
Cer_libreta de notas	Certificar	N/A	Opción múltiple			
Cer_M_50% logro	Certificar	Matemática		X		
Cer_M_plumeros	Certificar	Matemática	Opción múltiple			X
Cer_M_plumeros_b	Certificar	Matemática	Opción múltiple			X
Rec_C_Ciclo de Agua	Recoger	Ciencias	Opción múltiple			
Rec_C_Corazón	Recoger	Ciencias	Opción múltiple		X	
Rec_C_Fotosíntesis	Retroalimentar	Ciencias				
Rec_C_Marcelo	Recoger	Ciencias	Opciónmúltiple		X	
Rec_C_Traslación_3	Retroalimentar	Ciencias		X		
Rec_C_Tronco	Recoger	Ciencias		X		
Rec_C_Vertebrados	Recoger	Ciencias	Opción múltiple	X		
Rec_L_comprensión lectora	Recoger	Lenguaje	Opción múltiple		X	
Rec_L_Noticias	Recoger	Lenguaje	Opción múltiple			
Rec_L_Tarta1	Recoger	Lenguaje	Opción múltiple			
Rec_L_Tarta2	Recoger	Lenguaje	Opción múltiple		X	
Rec_M_Lápices	Recoger	Matemática	Opción múltiple			
Rec_M_Manzanas	Recoger	Matemática	Opción múltiple			
Retro_C_circuito	Retroalimentar	Ciencias	Opción múltiple			
Retro_C_Hongos	Retroalimentar	Ciencias	Opción múltiple		X	
Retro_C_Mamut	Retroalimentar	Ciencias	Opción múltiple			X
Retro_H_Plaza Italia	Retroalimentar	Historia	Opción múltiple			

Retro_L_Fábula	Retroalimentar	Lenguaje	Opción múltiple			
Retro_M_Gráfico1	Retroalimentar	Matemática	Opción múltiple			
Retro_M_Gráfico2	Retroalimentar	Matemática	Opción múltiple			X
Retro_M_Plumeros	Retroalimentar	Matemática	Opción múltiple		X	
Retro_M_Rectángulo	Retroalimentar	Matemática	Opción múltiple			
Retro_M_Resta	Retroalimentar	Matemática	Opción múltiple			

De este modo, el conjunto de decisiones tomadas tiene un impacto positivo en la confiabilidad del instrumento definitivo, siendo la confiabilidad original de las formas A y B cercana a 0,50, para terminar con un instrumento con un alfa de Cronbach de 0,61. Respecto de la validez del constructo, los ítems que conforman el instrumento definitivo responden a las 4 dimensiones teóricas definidas como constituyentes de la competencia evaluativa docente.

7.4. Cuestionario de prácticas

De las 37 preguntas incluidas en el pilotaje, tres de ellas fueron eliminadas por problemas en la construcción de las mismas, los cuales fueron detectados posteriormente. Por ello, los análisis se trabajaron con 34 preguntas totales.

Se revisó la distribución de las respuestas en cada pregunta, con el objetivo de descartar una concentración de las mismas en las opciones centrales o en algún extremo. Los datos muestran que existe variabilidad en los patrones de respuesta y que todas las opciones tienen un porcentaje razonable de selección, esto es, que cada opción presente al menos un 5% de selección. No fue posible descartar un efecto de la deseabilidad social en las respuestas de los docentes, pues no se cuenta con variables o casos de control para hacer una revisión.

En el AFE se extraen al menos 5 factores. Por lo que se procedió a realizar una selección de las preguntas que cargan en el primer factor, pues, al igual que en la prueba de competencias evaluativas, para trabajar con modelo IRT Rasch el requisito es que el instrumento mida una única dimensión subyacente o factor. Las cargas factoriales de los 25 reactivos del cuestionario que conforman el factor principal se presentan en la Tabla 8.

Cabe señalar, que la dimensión ‘analizar e interpretar evidencias de los aprendizajes’ fue excluida del instrumento, ya que las conductas observables que se pueden desprender de esta dimensión tienen alta deseabilidad bajo el formato de autorreporte. Se pilotearon 2 ítems que fueron validados en la etapa de jueces; sin embargo, aun cuando la evaluación no tenía consecuencias asociadas, estos ítems no se comportaron bien psicométricamente, por tanto, incluirlos inducía a conclusiones erróneas respecto de los resultados y se decidió eliminarlos. Si bien el número de ítems de la subdimensión recoger evidencias es menor al de las otras dos subdimensiones, son suficientemente robustos para poder entregar información confiable y válida.

La tabla de especificaciones del instrumento final es la siguiente:

Cuestionario	Recoger evidencia	Retroalimentar	Certificar
Nº ítems	6	10	9

Tabla 8. Pesos factoriales de los 25 ítems del cuestionario de prácticas evaluativas que cargan en un único factor

Dimensión	Posición	Instrucción	Reactivo	Peso factorial
Certificar	cuestN25	¿Con qué frecuencia realiza las siguientes prácticas al poner nota o calificar?	(-) Realizo autoevaluaciones con nota o puntaje.	-0,49
	cuestN28		(-) Quito u otorgo décimas o puntos por responsabilidad, participación esfuerzo.	-0,33
	cuestN29		(-) Pongo notas a algunas actividades para motivar que los estudiantes las realicen.	-0,23
	cuestN30		(+) En las preguntas abiertas de mis pruebas, incluyo los criterios con que serán corregidas.	0,49
	cuestN31		(+) Antes de poner nota a un trabajo, genero instancias para retroalimentar entregas parciales.	0,62
	cuestN32		(+) Si al corregir una prueba descubro una pregunta que formulé mal, no la considero en el puntaje final.	0,21
	cuestN34		(+) Explico anticipada y detalladamente a los estudiantes lo que se evaluará en las pruebas o trabajos con nota.	0,39
	cuestN36		(+) Después de corregir una prueba reviso con los estudiantes cada una de las preguntas con sus respuestas de correctas.	0,42
	cuestN37		(+) Analizo las notas finales para verificar si es que representan el nivel de aprendizaje alcanzado y las modifico si es necesario.	0,28
Recoger evidencia	cuestN04	Cuando hace preguntas o actividades en clase, para verificar si los estudiantes han comprendido lo que está enseñando, ¿Con qué frecuencia realiza cada una de las siguientes acciones?	(+) Pido a los estudiantes que dan una respuesta inesperada o incorrecta, que expliquen cómo llegaron a ella.	0,4
	cuestN05		(+) Pido a los estudiantes que respondieron lo que yo esperaba, que expliquen cómo llegaron a su respuesta.	0,43
	cuestN06		(+) Termino la clase haciendo una pregunta o ejercicio breve que todos deben responder por escrito y entregarme.	0,43
	cuestN08		(+) Recorro la sala y uso una pauta para verificar el nivel de comprensión que logran mis estudiantes durante el desarrollo de una actividad.	0,57
	cuestN09		(+) Planifico preguntas o ejercicios específicos para detectar si los estudiantes presentan alguna de las dificultades prototípicas de comprensión.	0,58
	cuestN10		(+) Uso estrategias para conocer la respuesta de todos o casi todos mis estudiantes (ej. que escriban su respuesta en un papel que recojo, o que voten por una respuesta).	0,52
Retroali-mentación formativa	cuestN12	Cuando un/a estudiante da una respuesta inesperada o incorrecta, u observa que no se ha logrado lo esperado en su trabajo, ¿con qué frecuencia hace lo siguiente?	(+) Le pido que explique cómo ha llegado a su respuesta.	0,57
	cuestN13		(-) Le indico dónde puede encontrar la respuesta correcta.	-0,25
	cuestN15		(+) Pregunto a otro(s) estudiante(s) si está(n) de acuerdo con la respuesta de su compañero y por qué.	0,24
	cuestN18	Cuando pide a los estudiantes hacer una actividad o elaborar un producto en clases, ¿con qué frecuencia realiza las siguientes acciones?	(+) Muestro modelos de trabajos bien logrados para ilustrar lo que espero de ellos.	0,39
	cuestN19		(+) Pido a los compañeros que intercambien sus trabajos y conversen sobre estos.	0,68
	cuestN20		(+) Pido a los compañeros que revisen mutuamente su trabajo usando una pauta de evaluación.	0,67
	cuestN21		(+) Les entrego una rúbrica o pauta para que vayan evaluando su propio trabajo a medida que lo hacen.	0,59
	cuestN22		(+) Si veo que algún estudiante no está logrando lo esperado, dialogo con él o ella sobre qué y cómo mejorar.	0,45
	cuestN23		(-) Si veo que algún estudiante no está logrando lo esperado, le repito las instrucciones por si no estaba atento.	-0,31
	cuestN24		(+) Si veo que algún estudiante no está logrando lo esperado, le indico específicamente qué debe hacer para mejorar su trabajo.	0,23

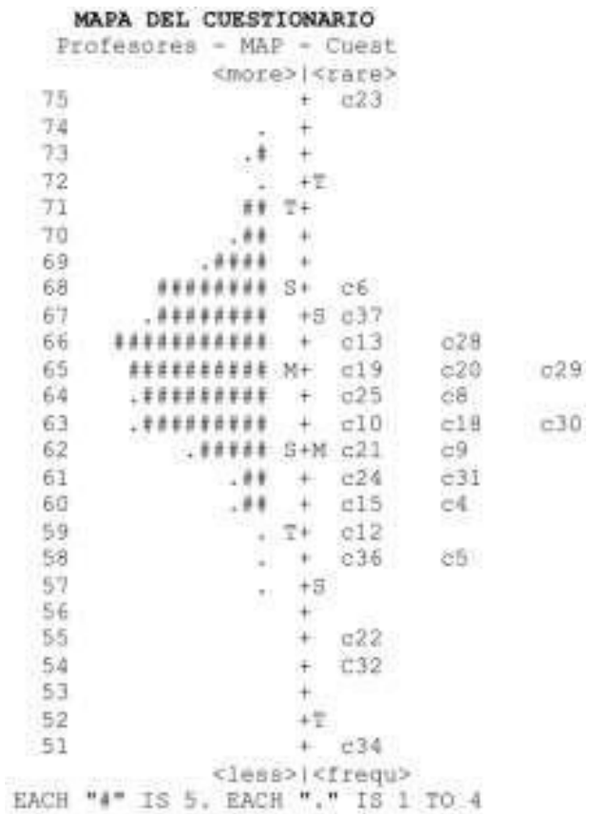
(-) La respuesta al ítem se codifica de manera inversa.
(+) La respuesta al ítem se codifica de manera directa.

7.4 Cuestionario definitivo de prácticas evaluativas

Los 25 reactivos seleccionados para construir el cuestionario definitivo de prácticas evaluativas obtienen una confiabilidad de 0,70.

En cuanto al análisis IRT, utilizando el modelo de respuesta graduada de Andrich, se construyó un mapa de los 25 reactivos y se calculó un puntaje en una escala de 25 a 100 puntos, manteniendo la equivalencia con la escala real. El mapa de reactivos se presenta en la Figura 11.

Figura 11. Mapa de ítems del cuestionario definitivo (25 ítems)



8. CONCLUSIONES

A continuación se presentan las principales conclusiones del estudio, ordenadas de acuerdo con los objetivos planteados en este.

Respecto del primer objetivo: diseñar instrumentos de autoaplicación online para diagnosticar el nivel de competencia en evaluación de aprendizajes que presentan los docentes que se desempeñan en enseñanza básica, se puede señalar que se elaboraron dos instrumentos (a) una prueba de selección múltiple y preguntas abiertas que integra el conocimiento pedagógico del contenido y habilidades de evaluación en contextos disciplinares específicos y (b) un cuestionario formulado bajo la forma de una escala de valoración tipo Likert que aborda la frecuencia con que el docente realiza determinadas prácticas. Estos dos formatos de instrumentos son coincidentes con las formas en que se evalúa en otros países la alfabetización evaluativa (Gotch y French, 2015). A diferencia de otros

instrumentos disponibles para estudiar alfabetización en evaluación de aprendizajes, en nuestra prueba se incorporaron situaciones de aula ancladas a asignaturas correspondientes a enseñanza básica, integrando la necesidad de dominar el contenido pedagógico de la disciplina.

La prueba aborda las cuatro dimensiones que se reconocen como las más relevantes para señalar que un docente tiene un dominio práctico de la evaluación (alfabetizado en evaluación), las cuales son: recoger evidencias de aprendizajes de sus estudiantes, analizar e interpretar evidencias de los aprendizajes, retroalimentar formativamente y certificar y calificar el aprendizaje de los estudiantes. Estas dimensiones cubren los cuatro dominios fundamentales de la competencia y son susceptibles de ser evaluados mediante este formato de preguntas y al utilizar casos se logra un nivel de complejidad cognitiva mayor que el solo reconocimiento. El instrumento en su composición actual privilegia las tareas en que el docente requiere retroalimentar el aprendizaje que tienen un peso algo mayor en la prueba. Si bien lo anterior fue el resultado del comportamiento psicométrico en la aplicación piloto, dada la relevancia de esta dimensión para promover el aprendizaje, consideramos que esta leve sobrerrepresentación de ningún modo distorsiona la validez del instrumento. Retroalimentar adecuadamente un desempeño, requiere analizarlo e interpretarlo acertadamente, por lo que el mayor número de tareas en este ámbito compensa, a nuestro juicio, apropiadamente la relativa menos presencia de tareas de analizar e interpretar la información.

En el cuestionario de prácticas evaluativas, se abordan tres dimensiones: recoger evidencias de aprendizajes de sus estudiantes, retroalimentar formativamente y certificar o calificar el aprendizaje de los estudiantes. La dimensión analizar e interpretar evidencias de los aprendizajes fue excluida, ya que las conductas que era posible explorar a través de un formato de autorreporte presentaban el riesgo de alta deseabilidad social, y no se comportaron bien psicométricamente. Como este es un cuestionario que recoge información sobre la frecuencia con que un docente realiza ciertas prácticas evaluativas específicas, y su resultado no se expresa en términos de nivel de dominio o competencia en evaluación, excluir la subdimensión de análisis e interpretación de información no tiene consecuencias relevantes para el docente que quiere conocer cómo son sus prácticas en relación con el grupo de referencia y a un referente teórico propuesto, ya que se explicita en el reporte de resultados qué ámbitos son abordados con este instrumento.

En relación con el segundo objetivo: realizar una aplicación piloto de los instrumentos para determinar sus características psicométricas. El análisis factorial exploratorio realizado con el conjunto de preguntas da cuenta de la agrupación de los ítems en dos factores, sin embargo, para hacer el análisis IRT y selección de preguntas, solo se seleccionó aquellas que cargaban en uno de los factores. Para las 24 preguntas con las que se hizo el análisis, se obtuvo un alfa de 0,64, una dificultad promedio clásica de 60% y una discriminación promedio de 0,37 y la selección para el instrumento definitivo presenta un alfa de Cronbach de 0,61 una dificultad IRT promedio de 24,88 y una discriminación punto biserial promedio de 0,33. La prueba final muestra una confiabilidad similar a la reportada por otros estudios (Gotch & French, 2015).

La prueba definitiva que cuenta con 20 ítems está disponible online y a ella se puede acceder on- line. Una vez finalizada la aplicación de la misma, entrega un reporte automático de un desempeño aproximado (indicando visualmente el rango de error), el cual tiene como cálculo a la base un puntaje IRT.

Por otra parte, el cuestionario tipo Likert sobre prácticas evaluativas, cuenta con 25 afirmaciones agrupadas en un solo factor y una consistencia interna con alfa de Cronbach de 0,7, similar a otros instrumentos de este tipo.

Sin embargo, el tercer objetivo, que consistía en describir niveles de logro de la competencia evaluativa de los docentes a partir de los resultados obtenidos en la aplicación piloto no se logró, debido a que el error de medición asociado al puntaje obtenido por los participantes fue alto y no se pudo establecer categorías de desempeño. Una posible explicación para esta baja consistencia de la prueba es que al ser dimensiones distintas, no hay un patrón de tendencias que vayan en el mismo sentido, esto es, una persona que tiene un alto desempeño en una de ellas no necesariamente tiene alto desempeño en las otras dimensiones o tareas y los patrones de respuesta de cada profesor no permiten generar un perfil consistente. El bajo nivel de conocimiento o alfabetización en el área de evaluación de aprendizajes en Chile y la disparidad en las dimensiones que habitualmente son más enfatizadas en la formación inicial (que prioriza la formulación de instrumentos por sobre las demás subdimensiones exploradas en este estudio) y continua (Agencia de la Calidad de la Educación, 2016) podría ayudar a explicar la escasa consistencia interna del instrumento.

La bibliografía especializada plantea que en el sistema escolar coexisten profesores formados bajo tres lógicas distintas que condicionan el enfoque que tienen en sus prácticas evaluativas: a) los formados en la década del 80 o anteriores, muy fuertes en dominio de construcción de instrumentos y análisis de datos cuantitativos pero sin ninguna formación en prácticas de retroalimentación ni de interacción pedagógica en aula; b) los formados en la década de los 90, cuyo conocimiento es menor en construcción de instrumentos y en estrategias de certificación y mayor en retroalimentación y evaluación formativa; y c) los formados después del 2000 que tienen un enfoque de complementariedad de evaluación sumativa y formativa y por tanto mayor dominio de las cuatro dimensiones evaluadas (Deneen & Brown, 2016). No obstante, esta clasificación no aplica del todo a Chile, ya que la mayoría de las instituciones formadoras de profesores se sitúan en la actualidad en alguno de los dos primeros enfoques y el tercero no se visibiliza en la formación inicial (Agencia de la Calidad de la Educación, 2016). De esta forma, entre los participantes se podrían presentar diversidad de dominios en las distintas dimensiones que hagan que la prueba pierda consistencia interna y no sea posible definir niveles para generar el mapa.

Por último, dado que la prueba actualmente disponible online está constituida por preguntas que se comportaron psicométricamente bien para formas distintas, se requiere volver a aplicarla para obtener una muestra de docentes que hayan respondido a estos ítems en su conjunto, y estudiar así sus características psicométricas. Por la confiabilidad que esta presenta, puede ser utilizada en contextos de investigación, donde se trabaje con datos agrupados que no implican consecuencias individuales. El reporte y puntaje que actualmente genera la plataforma está construido sobre la base de 13 de los 20 ítems disponibles online (cuya consistencia interna como escala fue estudiada y es reportada en este informe). Los 7 restantes se incluyeron para continuar recogiendo datos y realizar posteriormente un nuevo análisis del instrumento en su conjunto, situando todos los ítems en una misma escala, y de lograr un nivel adecuado de confiabilidad, elaborar niveles de logro. Esta, consideramos, una siguiente etapa del proyecto de investigación. En consecuencia, el resultado que hoy entrega el reporte debe tomarse en esta etapa con cautela y como una aproximación al nivel de dominio del docente en el ámbito de la evaluación de aprendizajes y por lo tanto, jamás utilizable en contexto de alguna consecuencia para este. Es necesario señalar, que al no ser aun posible reportar resultados expresados en niveles de logro, su contribución a un autodiagnóstico del docente es aún limitada, pues resulta difícil para quien lo rinde

concluir sobre necesidades específicas de formación. Sin embargo, durante el desarrollo del proyecto hubo al menos dos instancias de aplicación colectiva y presencial del instrumento, después de la cual se dio una discusión y reflexión sobre las propias prácticas muy interesante entre los docentes. Ellos reportaron que responder y luego discutir sobre cuál resultaba ser una mejor decisión pedagógica en cada tarea, les permitió revisar su práctica habitual y replanteársela. De esta manera, el instrumento en su versión actual puede resultar de utilidad para un diagnóstico en un sentido amplio, más que para una 'medición' de las propias habilidades en el terreno de la evaluación de aprendizajes.

En el caso del cuestionario de prácticas evaluativas, tiene un comportamiento psicométrico que permite su uso para el estudio sobre prácticas pedagógicas en los docentes chilenos, como también para la caracterización de las propias prácticas y obtención de un perfil individual en relación con el patrón promedio observado en la muestra para cada subdimensión evaluada, esto último es relevante ya que los docentes pueden hacer uso de él para conocer cómo están sus prácticas de evaluación y autogestionar instancias para ir aumentando o disminuyendo su frecuencia (según sea lo deseable teóricamente) y monitorear si presentan un cambio en el tiempo, por ejemplo.

9. RECOMENDACIONES DE POLÍTICA PÚBLICA

El análisis psicométrico de la aplicación piloto del Cuestionario de Prácticas de Evaluación de Aula permitió seleccionar un conjunto de reactivos para conformar un instrumento válido y con un nivel de confiabilidad suficiente para estudiar dichas prácticas a través del reporte de los docentes sobre ellas. Al mismo tiempo, el escalamiento de los reactivos, usando el modelo de Rasch, y el análisis de su ordenamiento en un 'mapa de ítems' permite anticipar que dicho instrumento puede aportar a la descripción de un continuo coherente de niveles crecientes de desempeño en evaluación de aprendizajes. Así por ejemplo, en el ámbito de la retroalimentación, se ubicaron en los niveles más bajos del mapa de ítems aquellos que reportan prácticas básicas como dar retroalimentación grupal después de calificar una prueba ('después de corregir una prueba reviso con los estudiantes cada una de las preguntas con sus respuestas correctas'), mientras que aquéllos en que se da mayor protagonismo a los estudiantes proporcionando una guía para la evaluación entre pares, se ubicaban en los niveles superiores ('Pido a los compañeros que revisen mutuamente su trabajo usando una pauta de evaluación'). Algo semejante sucede con los reactivos referidos a las dimensiones de 'calificar los aprendizajes' y 'recoger evidencia'. En consecuencia, este instrumento queda a disposición para estudiar prácticas de evaluación de los docentes en Chile, lo que permitirá profundizar el diagnóstico sobre las prácticas de evaluación de aula en el país, a la vez que generar una herramienta que favorezca la reflexión de los docentes sobre ellas.

Este cuestionario también puede servir como un instrumento que complemente la información obtenida con otros medios como observación de clases, con lo que se podría estudiar la brecha que hay entre prácticas autorreportadas y prácticas observadas y comparar los resultados de aprendizajes de los estudiantes en profesores con distintos perfiles de prácticas reportados.

Respecto de la prueba, se encuentra disponible un instrumento on-line con un reporte automático de resultados obtenidos, el que puede ser usado como:

- Diagnóstico general de dominio sobre evaluación de aprendizajes. Si bien su reporte no entrega un nivel de desempeño preciso, permitirá a los docentes tener información sobre una posición de rendimiento (de mayor o menor cercanía al máximo puntaje de la prueba) y a partir de este resultado, definir sus necesidades de formación continua.

- Realización de talleres donde se responda el instrumento y luego se comenten los ítems de manera individual o se analice con los docentes las respuestas dadas y las razones por las que marcaron esa opción y no alguna de las otras. Esta dinámica fue probada en un taller con un grupo de profesores durante la aplicación piloto y resultó muy motivante y reveladora para los docentes.

Entre las proyecciones para la prueba, está realizar estudios relacionando los resultados obtenidos con esta prueba y el desempeño en las dimensiones de la evaluación docente asociadas a evaluación para obtener evidencia sobre su validez. Este estudio de validez puede servir también como línea de base para analizar el progreso en el dominio evaluativo de los docentes a partir de las políticas de refuerzo de la evaluación de aula que el Ministerio de educación a través del CPEIP y los lineamientos curriculares, así como las políticas implementadas por la Agencia de la Calidad de la Educación enfocadas a potenciar las prácticas de evaluación formativa en el aula.

REFERENCIAS

Agencia de la Calidad (2016) Estudio sobre formación inicial docente en evaluación educacional, Santiago de Chile.

Bambrick-Santoyo, P. (2010). *Driven by Data: A Practical Guide to Improve Instruction*. San Francisco, California: Jossey-Bass.

Black, P., & Wiliam, D. (1998). *Inside the black box: raising standards through classroom assessment*. Granada: Learning.

Black, P. J., & Wiliam, D. (2004). The formative purpose: assessment must first promote learning. In Wilson, M. (Ed.). *Towards coherence between classroom assessment and accountability*. 3rd Yearbook of the National Society for the Study of Education (part. 2) (vol. Part II, pp. 20-50). Chicago, IL: University of Chicago Press.

Black, P., Wilson, M. & Yao, S. (2011). Road Maps for Learning: A Guide to the Navigation of Learning Progressions. *Measurement: Interdisciplinary Research and Perspectives 9(2-3)*, 71-123.

Bocala, C., & Boudett, K. P. (2015). Teaching educators habits of mind for using data wisely. *Teachers College Record*, 117(4),1-20.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education*, *38*(6), 698-712.

Brookhart, S. M. (2008). *How to give effective feedback to your students*. Alexandria, VA: Association for Supervision and Curriculum Development.

Brookhart, S. M. (2012). Grading. En McMillan, J. H. (Ed.). (2012). *SAGE Handbook of Research on Classroom Assessment*. SAGE Publications. Sage Publications.

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

Carr, M., McGee, C., Jones, A., McKinley, E., Bell, B., Barr, H. & Simpson, T. (2000). *Strategic research initiatives: the effects of curricula and assessment on pedagogical approaches and on educational outcomes*. Wellington, New Zealand: Ministry of Education.

Castillo, S. & Cabrerizo, J. (2003). *Evaluación educativa y promoción escolar*. Madrid: Pearson Educación.

Celman, S. (2005) ¿Es posible mejorar la evaluación y trasformarla en una herramienta de conocimiento? En A. Camilloni et al. (Comp.). *La evaluación de los aprendizajes en el debate didáctico contemporáneo* (pp. 35-66). Buenos Aires: Paidós Educador.

Chin, C. (2007). Teacher questioning in science classrooms: Approaches that stimulate productive thinking. Journal of research in Science Teaching, 44(6), 815-843.

Costello, A., & Osborne, J. (2005). Exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(7), 1–9.

Covacevich, C. (2014). *Cómo seleccionar un instrumento para evaluar aprendizajes estudiantiles*. Nota Técnica; 738; Banco Interamericano de Desarrollo.

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. Applied measurement in Education, 12(1), 53-72.

Danielson, C. (2013). The Framework for Teaching Evaluation Instrument. Princeton, NJ: The Danielson Group.

Datnow, A., & Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decisión making: A literature review of international research. Journal of Educational Change, 17(1), 7-28.

Deneen, C.C. y Brown, G. (2016). The impact of conceptions of assessment on assessment literacy in a teacher education program. Cogent Education, 3: 1225380; 1-14.

Dewey, J. (1928, March). Progressive education and the science of education. Paper presented at the eighth annual conference of the Progressive Education Association, Washington, D.C.

Earl, L., (2003). *Assessment as Learning: Using classroom assessment to maximise student learning*. Thousand Oaks, CA: Corwin Press.

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.).

Equipo de Tarea para la Revisión del SIMCE. (2015). Hacia un sistema completo y equilibrado de evaluación de los aprendizajes en Chile. Recuperado de http://www.mineduc.cl/wp-content/uploads/sites/19/2015/11/Informe-Equipo-de-Tarea-Revisi%C3%B3n-Simce.pdf

Förster C., & Rojas-Barahona, C. (2008). Evaluación al interior del aula: una mirada desde la validez, confiabilidad y objetividad. *Rev. Pensamiento Educativo*, 43, 285-305.

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

Gimeno, J. (2010). ¿Qué significa el currículum? En Gimeno Sacristán, J. (Ed.), *Saberes e incertidumbres del currículum* (pp.19-44). Madrid: Morata.

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice*, *33*(2), 14-18.

Goubeaud, K. (2010). How is Science Learning Assessed at the Postsecondary Level? Assessment and Grading Practices in College Biology, Chemistry and Physics. *Journal of Science Education and Technology, 19(3*), 237-245.

Griffin, P. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation, 33(1)*, 87-99.

Harlen, W. (2007). Formative classroom assessment in science and mathematics. *Formative classroom assessment: Theory into practice,* 116-135.

Hattie, J. & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77(1)*, 81– 112.

Hattie, J. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. London: Routledge.

Heritage, M., & Heritage, J. (2011). Teacher questioning: The epicenter of instruction and assessment. *Applied Measurement in Education, 26(3*), 176-190.

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

Hotiu, A. (2006). *The relationship between item difficulty and discrimination indices in multiple-choice tests in a Physical science course* (MSc thesis). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.599.5172&rep=rep1&type=pdf

Hutchison, D., Francis, M., & Griffin, P. (2014). Developmental teaching and assessment. En P. Griffin (Ed.), *Assessment for teaching* (p. 26–57). New York, NY: Cambridge University Press.

Ingram, D., Louis, K. S., & Schroeder, R. G. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record, 106(6),* 1258-1287.

Jensen, J.L., McDaniel, M.A., Woodard, S.M. & Kummer, T.A. (2014). Teaching to the Test…or Testing to Teach: Exams Requiring Higher Order Thinking Skills Encourage Greater Conceptual Understanding. *Educ Psychol Rev, 26(2)*; 307-329. https://doi.org/10.1007/s10648-013-9248-9

Joint Committee on Standards for Educational Evaluation. (2003). *The Student Evaluation Standards: How to Improve Evaluations of Students*. Thousand Oaks, CA: Corwin Press, Inc.

Jones, A. & Moreland, J. (2005). The importance of pedagogical content knowledge in assessment for learning practices: a case-study of a whole-school approach. *The Curriculum Journal, 16*(2), 193- 206.

Kluger, A., De Nisi, A. (2000) The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin, Vol. 119(2), Mar 1996, 254-284.

Kahl, S. R., Hofman, P., & Bryant, S. (2012). Assessment literacy standards and performance measures for teacher candidates and practicing teachers. *Assessment*.

Laurillard, D. (2002). *Rethinking university teaching: a conversational framework for the effective use of learning technologies.* (2nd ed.). London: Routledge.

Leahy, C. Lyon, M. Thompson, M., & Wiliam, D. (2005). Classroom Assessment: Minute by Minute, Day by Day. *Journal of the Department of Supervision and Curriculum Development, N.E.A, 63(3),* 19-24.

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

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. In Lederman (Eds.) *Examining pedagogical content knowledge* (pp. 95-132). Dordrecht: Springer.

Mandinach, E. & Gummer, E. (2012). *Navigating the landscape of data literacy: It IS complex.* Washington, DC & Portland, OR: WestEd and Education Northwest.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20(1),* 20-32.

McMunn, N., Schenck, P., & McColskey, W. (2003). Standards-Based Assessment, Grading, and Reporting in Classrooms: Can District Training and Support Change Teacher Practice? Paper presented at the annual meeting of AERA, Chicago.

Mercado, A. & Martínez, F. (2014). Evidencias de prácticas de evaluación de un grupo de profesores de primarias de Nuevo León. *Revista Mexicana de Investigación Educativa, 19(61),* 537-567.Recuperado dehttp://www.scielo.org.mx/pdf/rmie/v19n61/v19n61a9.pdf

Mineduc. (2008). Marco para la buena enseñanza. Santiago: Autor.

Mineduc, 2015. Resultados de la Evaluación Docente 2014. Recuperado en http://www.docentemas.cl/docs/Resultados_Evaluacion_Docente_2014.pdf

Mineduc. (2016). Plan de Aseguramiento de la Calidad de la Educación, 2016-2019. Santiago: Autor.

Ministerio de Educación (2017) Criterios y Normas Mínimas Nacionales sobre Evaluación, Calificación y Promoción Escolar de estudiantes de Educación Regular en sus niveles Básico y Medio formación General y Diferenciada, documento presentado al Consejo Nacional de Educación.

Ministerio de Educación. (2012). Estándares orientadores para carreras de pedagogía en educación media. Santiago: Chile: Autor.

Ministerio de Educación. (2014). Hacia un Sistema completo y equilibrado de evaluación de aprendizajes en Chile. Informe del Equipo de Tarea para la Revisión del Sistema del SIMCE.

Mishra, P. & Koehler, M. (2006). Technological pedagogical content knowledge: a framework for teacher knowledge. *Teachers College Record*, *108*(6), 1017-1054.

OCDE (2005) Formative Assessment: Improving Learning in Secondary Classrooms, Policy brief, November 2005.

Panadero, E., & Brown, G. T. (2017). Teachers' reasons for using peer assessment: positive experience predicts use. European Journal of Psychology of Education, 32(1), 133-156.

Panadero, E., Alonso-Tapia, J., & Huertas, J.A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences, 22,* 806-813. doi: 10.1016/j.lindif.2012.04.007

Popham, W. J. (2004). Why assessment illiteracy is professional suicide. *Educational Leadership, 62,* 82 -83.

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *Teacher Educator, 46,* 265-273.

Popham, W. J. (2013). *Classroom assessment: What teachers need to know* (7th ed.). Boston: Pearson.

Prieto, M. & Contreras, G. (2008). Las concepciones que orientan las prácticas evaluativas de los profesores: un problema a develar. *Estudios Pedagógicos, 34(2),* 245-262.

Race, P., Brown, S., & Smith, B. (2005). *500 Tips on Assessment*. Second edition. USA, NY: RoutledgeFalmer.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education, 26(7),* 1372-1380.

Ravela, P. (2009). Consignas, devoluciones y calificaciones: los problemas de la evaluación en las aulas de educación primaria en América Latina. *Páginas de Educación, 2*, 49-89.

Ravela, P., Leymonié, J., Viñas, J., y Haretche, C. (2014). La evaluación en las aulas de secundaria básica en cuatro países de América Latina. *Propuesta educativa, 41(1)*, 20-45.

Ruiz-Primo, M. A. (2011). Informal formative assessment: The role of instructional dialogues in assessing students' learning. *Studies in Educational Evaluation, 37*, 15–24. doi: 10.1016/j.stueduc.2011.04.003

Ruiz-Primo, M. A., & Li, M. (2013). Analysing Teachers' Feedback Practices in Response to Students' Work in Science Classrooms. Applied Measurement in Education, 26(3), 163–175. doi: 10.1080/08957347.2013.793188

Sanmartí, N. (2007). *10 ideas clave: evaluar para aprender*. España: Graó.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15(2)*, 4- 31.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of a new reform. Harvard Educational Review, 57 (1), 1-22.

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research; 78(1)*, 153-189. doi: 10.3102/0034654307313795

Stiggins, R. (2002). Assessment crisis: the absence of assessment for learning. *Phi Delta Kappan, 83,* 758-765.

Stiggins, R. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan, 86*, 22 -27.

Sun, Y., Correa, M., Zapata, A. & Carrasco, D. (2011). Resultados: qué dice la evaluación docente acerca de la enseñanza en Chile. En J. Manzi, R. González & Y. Sun (Eds.), *La evaluación docente en Chile* (pp. 91-135). Santiago, Chile: Pontificia Universidad Católica de Chile, Centro de Medición MIDE UC.

Tejedor, F., & García-Valcárcel, A. (2010). Evaluación del desempeño docente. *Revista Española de Pedagogía, 68(247)*, 439-459.

Topping, K. (1998). Peer assessment between students in colleges and universities. Review of educational Research, 68(3), 249-276.

Torres, M. & Cárdenas, E. (2010). ¿Qué y cómo se ha investigado sobre la evaluación de los aprendizajes en los últimos cinco años? Estado del arte de las investigaciones (2005-2010). *Enunciación, 15(1*), 141-156. Recuperado de http://dialnet.unirioja.es/descarga/articulo/3661659.pdf

Tunstall, P., & Gipps, C. (1996). Teacher Feedback to Young Children in Formative Assessment: A Typology. *British Educational Research Journal, 22(4)*. 389-404.

Villardón, L. (2006). Evaluación del aprendizaje para promover el desarrollo de competencias. Educatio. *Siglo XXI, 24*, 15-35.

Vinas-Forcade, J. y Emery, C. (2015). *Las consignas de pruebas escritas como herramienta de evaluación del desempeño docente*. Recuperado de http://www.colmee.mx/public/conferences/1/presentaciones/ponenciasdia3/42Las_consignas. pdf

Wiliam, D. (2011). What is assessment for learning? *Studies in educational evaluation, 37(1)*, 3-14. 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(1),* 49-65.

Zlokovich, M. (2001). Teaching Tips. Grading for Optimal Student Learning. *APS Observer 14(1)*. Recuperado de:http://www.psychologicalscience.org/teaching/tips/tips_0101.cfm

Zohar, A. & Schwartzer, N. (2005). Assessing teachers' pedagogical knowledge in the context of teaching higher-order thinking. *Journal of Science Education, 27(13)*, 1595-1620.