

Evaluación de la escritura a gran escala: análisis de la prueba de escritura del Tercer Estudio Regional Comparativo y Explicativo (TERCE)

ANA ATORRESI*

Universidad Nacional de Río Negro – Argentina

FRANK VILLEGAS**

Pontificia Universidad Católica del Perú – Perú

CARLOS PARDO ADAMES***

Universidad Católica de Colombia – Colombia

Recibido el 24-08-18; primera evaluación el 12-08-19;
segunda evaluación el 13-08-19; tercera evaluación el
26-08-19; aceptado el 27-08-19

RESUMEN

La prueba de escritura del Tercer Estudio Comparativo y Explicativo, implementada por el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación, constituye una de las pocas experiencias internacionales de evaluación a gran escala de la escritura. Para fortalecer estas experiencias, se revisan

* Doctoranda en Ciencias Sociales (Universidad de Buenos Aires). Especialista en Constructivismo y Educación (Facultad Latinoamericana de Ciencias Sociales) y Profesora de Castellano, Literatura y Latín (Instituto Nacional Superior del Profesorado Secundario). Profesora asociada regular de Prácticas de Enseñanza en el Profesorado en Lengua y Literatura e investigadora en el Centro de Estudios del Lenguaje, la Literatura, su Aprendizaje y su Enseñanza (Universidad Nacional de Río Negro). Fue docente de grado y posgrado en la UBA, la UNQ y la FLACSO y profesora invitada en la Universidad Nacional de Colombia. Investiga el aprendizaje, la enseñanza y la evaluación de la lectura y la escritura. Correo electrónico: aatorresi@unrn.edu.ar

** Magíster en Cognición, Aprendizaje y Desarrollo. Licenciado en Educación Primaria por la Pontificia Universidad Católica del Perú (PUCP). Actualmente, se desempeña como docente en la Facultad de Educación y la Escuela de Posgrado de la PUCP. Sus temas de interés y especialización son cognición docente, didáctica de la comunicación y evaluación educativa. Correo electrónico: fvillegas@pucp.edu.pe

** Doctorando en Psicología por la Universidad Católica de Colombia. Magíster en Educación por la Pontificia Universidad Javeriana. Psicólogo por la Universidad Nacional de Colombia. Docente de grado y posgrado en temas relacionados con evaluación, formación de competencias y psicometría por la Universidad Católica de Colombia y la Universidad Sergio Arboleda. Investigador de Colciencias. Correo electrónico: capardo@ucatolica.edu.co



el marco teórico adoptado en dicha prueba, las consignas aplicadas, la rúbrica y el proceso de codificación, combinando técnicas cualitativas y cuantitativas. Los resultados evidencian problemas vinculados con la validez y la confiabilidad, y permiten formular recomendaciones para futuras evaluaciones internacionales de la escritura. Estas comprenden la articulación entre el marco teórico y los objetivos de la evaluación, la revisión de los temas y la sintaxis de las consignas, y el empleo de estrategias para obtener evidencias de validez y confiabilidad en las distintas etapas del proceso evaluativo.

Palabras clave: prueba de escritura, análisis de tareas, rúbricas (guías de puntuación), validez, confiabilidad

Large-scale writing assessment: analysis of the Third Regional Comparative and Explanatory Study (TERCE) writing test

ABSTRACT

The writing test of the Third Comparative and Explanatory Study implemented by the Latin American Laboratory for Assessment of the Quality of Education constitutes one of the few international experiences of large-scale assessment of writing. In order to strengthen these experiences, the theoretical framework adopted in this test, the task applied, the rubric and the coding process are reviewed, combining qualitative and quantitative techniques. The results show problems related to validity and reliability, and allow to formulate recommendations for future international assessments of writing. These include an articulation between the theoretical framework, the objectives and the scope of the assessment, a process of thorough review of the topics and syntax of the prompts, and the use of strategies to obtain evidences of validity and reliability at the different stages of the assessment process.

Keywords: writing test, task analysis, rubrics (scoring guides), validity, reliability

Avaliação da escrita em larga escala: Análise da prova de escrita do Terceiro Estudo Comparativo e Explicativo (TERCE)

RESUMO

A prova de escrita do Terceiro Estudo Comparativo e Explicativo implementado pelo Laboratório Latino-Americano de Avaliação da Qualidade da Educação constitui uma das poucas experiências internacionais de avaliação em larga escala da escrita. Para fortalecer essas experiências, são revisados o marco teórico adotado na prova mencionada, as tarefas aplicadas, a rubrica e o processo de codificação, combinando técnicas qualitativas e quantitativas. Os resultados mostram problemas relacionados a validade e confiabilidade e permitem formular recomendações para futuras avaliações internacionais da escrita. Estas compreendem a articulação entre o marco teórico, os objetivos e os alcances da avaliação, a revisão dos tópicos e

a sintaxe das tarefas, e o uso de estratégias para obter evidências de validade e confiabilidade nas diferentes etapas do processo de avaliação.

Palavras-chave: prova de escrita, análise de tarefas, rubricas (guías de pontuação), validade, confiabilidade

1. INTRODUCCIÓN

Este trabajo busca caracterizar, desde un punto de vista externo, algunos aspectos de la evaluación de la escritura de los estudiantes latinoamericanos de tercer y sexto grado emprendida por el Tercer Estudio Regional Comparativo y Explicativo (TERCE) del Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE), de la Oficina Regional para América Latina y el Caribe (OREALC/Unesco)¹. La indagación de la escritura, por parte del LLECE, es una de las pocas iniciativas emprendidas por programas internacionales de evaluación de la calidad educativa dirigidas a estudiar las habilidades de producción textual. El LLECE ha realizado dos evaluaciones de la escritura: una en su Segundo Estudio (SERCE), cuyos resultados se publicaron en 2010 (Atorresi y cols., 2010) y la otra en el TERCE, con resultados dados a conocer cinco años después (Flotts, Manzi, Jiménez, Abarzúa, Cayuman y García, 2015)².

Es necesario destacar que la OREALC emprenda evaluaciones del desempeño en escritura de los estudiantes latinoamericanos de nivel primario, pese a los desafíos que supone evaluar la escritura a gran escala. La importancia de estas evaluaciones —así como de aquellas implementadas por sistemas nacionales de medición, como los de Colombia, Chile, México y Perú— radica en que la escritura es central como práctica social y habilidad humana cuyo aprendizaje comienza antes del ingreso en la escuela (Tolchinsky, 2017), se extiende a lo largo de toda la vida (Bazerman, 2017) y excede los currículos y usos escolares de la lengua escrita (Márquez, Iparraguirre y Bengtsson, 2015).

El trabajo analiza la prueba de escritura del TERCE, en cuanto al marco teórico adoptado, las consignas dadas a los estudiantes, la rúbrica elaborada para analizar los escritos y el proceso de codificación, combinando técnicas cualitativas y cuantitativas usuales en la evaluación educativa. Así, busca alentar los relativamente escasos análisis de la gran cantidad de información producida por los estudios evaluativos y puesta a disposición de investigadores (Murillo y Martínez-Garrido, 2016).

¹ El trabajo forma parte de una investigación más amplia, emprendida por los autores en el marco de un proyecto financiado por [eliminado para preservar el anonimato].

² El Primer Estudio o PERCE no analizó la escritura de los estudiantes.

2. MARCO TEÓRICO

2.1. Tipos de aproximación a la escritura, validez y confiabilidad

Dos aproximaciones han predominado en la evaluación de la escritura a gran escala: la directa y la indirecta (Huot, 2002). La primera supone que la evaluación debe basarse en el desempeño del estudiante al construir un texto escrito, por lo que se realiza con pruebas que solicitan producir textos completos e implica la participación de codificadores; la indirecta se realiza con pruebas de selección de respuesta, lo cual, para algunos investigadores, tiene la ventaja de excluir la subjetividad de los codificadores, y reducir el tiempo y costo que demanda el análisis de textos escritos (Bejar, 2017).

El concepto de validez refiere al grado en que la evidencia empírica y la teoría apoyan la interpretación y el uso de los resultados de una prueba (AERA, APA y NCME, 2014). Usualmente, las evaluaciones nacionales e internacionales reportan dos tipos de evidencia de validez: la basada en el contenido y en la estructura interna de una prueba. La evidencia basada en el contenido supone que los elementos de una prueba representan el *constructo* que se pretende evaluar (AERA, APA y NCME, 2014) y se obtiene mediante el juicio de expertos (Kane, 2016). La evidencia basada en la estructura interna de una prueba refiere a la consistencia entre la estructura empírica —la relación entre los ítems— y lo que la teoría indica sobre aquello que se evalúa (AERA, APA y NCME, 2014); para obtenerla, se utilizan técnicas estadísticas como el análisis factorial o el análisis de componentes principales. La primera técnica, por ejemplo, se usó en la prueba de lectura del TERCE (Flotts *et al.*, 2016b) y la segunda, en las pruebas de escritura de México y Perú (Instituto Nacional de para la Evaluación de la Educación [INEE], 2013; Ministerio de Educación [Minedu], 2016). Estas técnicas suelen ser utilizadas para obtener evidencias de unidimensionalidad, es decir, la predominancia de un solo constructo en la base del conjunto de ítems de la prueba (Burga, 2006). La noción de validez ha sido clave al optar por una aproximación directa o indirecta: diversos autores han señalado que la aproximación indirecta sobresimplifica y trivializa el amplio rango de habilidades y conocimientos involucrados en la escritura (Greemberg, 1992; Bridgeman, 2016).

Por su parte, el concepto de confiabilidad expresa el grado de consistencia de los resultados de una prueba en diversas aplicaciones (AERA, APA y NCME, 2014). En evaluaciones de la escritura con aproximación directa, suele reportarse el índice de consistencia entre los evaluadores, como se observa en el reporte técnico de la prueba nacional de escritura de Chile de 2013 (Agencia

de Calidad de la Educación [ACE], 2015). En algunos casos, se reportan, además, índices de consistencia interna; por ejemplo, en la prueba de escritura de México, se reporta el índice Alfa de Cronbach (INEE, 2013) y en la de Perú, el índice de separación de personas (Minedu, 2016).

2.2. Consignas y evaluación de la escritura

Se denominan *consignas* a las pautas que se brindan a los estudiantes para realizar una tarea (producir un texto escrito, en este caso). Diversos autores han planteado recomendaciones para la construcción de consignas de escritura, como describir una situación comunicativa, especificar las características del texto (género, tipo de secuencia, propósito, audiencia y extensión) e indicar el tiempo del que se dispone para realizar la tarea (Calfee y Miller, 2007; Donovan y Smolkin, 2017; Li, 2014; Ruiz-Funes, 2015). Asimismo, la formulación de la consigna debe ser simple y estar libre de ambigüedades, a fin de que su interpretación no se convierta en un obstáculo y de obtener, dentro de lo posible, escritos comparables entre sí; además, de acuerdo con lo que se busque evaluar, al elaborar la consigna, debe procurarse anticipar el grado de conocimiento previo del escritor sobre el tema (Atorresi, 2005). Cada uno de los componentes de una consigna incide en la escritura, razón por la cual se recomienda pilotear una serie amplia de consignas para decidir cuáles incluir en una evaluación (Bachman, 2002).

2.3. Rúbricas y codificación de textos escritos

Una rúbrica contiene los aspectos que deben considerarse al evaluar y gradaciones de calidad de esos aspectos (Wiggins, 1998). En la evaluación de la escritura con aproximación directa, las rúbricas permiten analizar los textos producidos por los estudiantes. Existen dos tipos de rúbricas: holísticas y analíticas. Las holísticas permiten obtener una evaluación cualitativa general, mediante una aproximación global y rápida a lo evaluado; las analíticas, en cambio, desagregan los criterios y exigen asignar un nivel o puntuación a cada uno de ellos, mediante una aproximación parsimoniosa a lo evaluado que resulta apropiada para caracterizar pormenorizadamente componentes textuales o habilidades específicas (Chan, Inoue y Taylor, 2015). Por ejemplo, en la prueba nacional de escritura de Chile del año 2013 (ACE, 2015), se empleó una rúbrica holística, mientras que en las pruebas nacionales de Perú y México (INEE, 2013; MINEDU, 2016; Villegas, 2016) y en las pruebas de escritura del TERCE (Flotts et al., 2015), rúbricas analíticas.

Por ser parte constitutiva de la evaluación de la escritura, las rúbricas deberían ser validadas. No obstante, revisiones recientes revelan que ofrecer evidencias sobre la validez de las rúbricas es relativamente poco frecuente (Jonsson y Svingby, 2007).

Diversos estudios advierten dificultades en el análisis de los textos escritos, como diferencias de interpretación de las rúbricas entre los codificadores según su formación previa y experiencia, y el modo y plazo en los que fueron formados; sesgos de indulgencia o tendencia a dejar pasar errores y de severidad o tendencia a ver más errores de los existentes; falta de consistencia del codificador a lo largo del proceso de codificación o privilegio de la variedad lingüística dominante en desmedro de otras variedades (Bejar, 2017; Behizadeh, 2014; Mendoza, 2018). Para asegurar la calidad de la codificación, usualmente se procura controlar el grado de acuerdo entre los codificadores. Para ello, la mayoría de los sistemas evaluación proporcionan entrenamiento (Johnson, Penny y Gordon, 2009).

Actualmente, se extiende el uso del modelo Rasch de múltiples facetas que permite estimar el grado de acuerdo entre codificadores, su severidad o indulgencia, así como la consistencia del codificador consigo mismo, además de características de los ítems y los evaluados (Eckes, 2015).

Aunque el proceso de codificación suele discutirse en términos de confiabilidad, también puede asociarse con la validez. Como los codificadores median la interpretación del desempeño en relación con el constructo, la evidencia de la calidad de la codificación provee información sobre el grado en que las puntuaciones —producto de la codificación— pueden ser interpretadas y usadas con fines específicos (Wind y Peterson, 2017).

3. METODOLOGÍA

Se realiza un análisis que combina técnicas cualitativas y cuantitativas. El análisis cualitativo consiste en el juicio de experto (Kane, 2016) del marco teórico, las consignas y la codificación a partir del examen de la totalidad de los documentos publicados por la OREALC/Unesco sobre la prueba de escritura:

- Unesco y MIDE UC (2011). *Tercer Estudio Regional Comparativo y Explicativo (TERCE). Manual para la elaboración de ítems.* (s/d).
- Flotts, M. P., Manzi, J., Jiménez, D., Abarzúa, A., Cayuman, C. y García, M.J. (2015). Los aprendizajes en escritura. *Tercer Estudio Regional Comparativo y Explicativo (TERCE). Informe de resultados. Logros de aprendizaje;* pp. 95-145. Santiago: OREALC/Unesco.

- Flotts, M. P., Manzi, J., Lobato, P., Durán, M. I., Díaz, M. P. y Abarzúa, A. (2016a). *Aportes para la enseñanza de la escritura*. Santiago: OREALC/Unesco.
- Flotts *et al.* (2016b). *Tercer Estudio Regional Comparativo y Explicativo (TERCE). Reporte técnico*. Santiago: OREALC/Unesco.

El análisis cuantitativo consiste en un análisis psicométrico de las rúbricas empleadas en la instancia de evaluación definitiva, mediante el modelo Rasch de créditos parciales, el cual permite considerar distintos niveles de logro para un mismo indicador (Bond y Fox, 2007). El *Reporte técnico* del TERCE no informa un análisis de este tipo para la prueba de escritura (cfr. Flotts *et al.*, 2016b, pp.261-264). Tampoco presenta información sobre los codificadores, razón por la cual no es posible emplear aquí un modelo que estime la habilidad de los evaluados calculando, además, el efecto de los codificadores.

La base de datos inicial de escritura contiene un total de 63 884 y 64 898 casos en tercer y sexto grado, respectivamente. Para este estudio, se eliminó el grupo de casos con puntuación máxima o mínima en todos los indicadores de la rúbrica, dado que su inclusión distorsiona otras estimaciones psicométricas y se evitó truncar los resultados en la escala (a diferencia del TERCE en las pruebas de lectura, matemática y ciencias). Esta eliminación corresponde al 4,7% de los casos de tercer grado y al 4,2% en sexto, por lo que se tomó un total de respuestas de 51 683 de tercer grado y 58 316 de sexto.

4. RESULTADOS

4.1. Sobre el marco teórico

La prueba de escritura del TERCE, de alcance latinoamericano, es una de las pocas evaluaciones internacionales que estudia la escritura de los estudiantes de primaria con aproximación directa. En este sentido, cumple con la condición más amplia de validez de contenido, por considerar que la escritura debe evaluarse a través de la producción escrita. No obstante, en un plano más específico, se aprecian diferencias entre el modo en que la escritura es evaluada y el modo en que resulta caracterizada en *Aportes para la enseñanza de la escritura* (*Aportes*, de aquí en adelante).

En *Aportes*, se reseñan diversas perspectivas disciplinares enfocadas en la escritura. Por un lado, se presenta el modelo de Bereiter y Scardamalia (1992), el cual describe los procesos mentales y modos de operar con el conocimiento implicados en la composición escrita de sujetos más y menos expertos.

Por otro lado, se referencia la propuesta sociocultural de Barton y Hamilton (2000), la cual postula que las prácticas de escritura suponen ideologías que norman el uso de los textos y prescriben el acceso a ellos, y que no pueden ser observadas directamente. Asimismo, para justificar las dimensiones escogidas en las rúbricas, *Aportes* se vale de la conceptualización del discurso académico que, con miras a su enseñanza, hacen Snow y Uccelli (2009).

No obstante, de la exposición del modelo de Bereiter y Scardamalia (1992), no se deriva una propuesta de evaluación que atienda a los procesos mentales involucrados en la composición escrita y al modo en que los menos expertos (presumiblemente los estudiantes de los grados evaluados) operan con el conocimiento al escribir. Al contrario, se afirma que «dadas las posibilidades de los instrumentos de evaluación de la escritura en TERCE, este estudio se hizo cargo *solo del producto final* del proceso de composición de estudiantes de tercer y sexto de primaria» (Flotts et al., 2016a, p. 12). En segundo lugar, con respecto a la revisión del planteo de Barton y Hamilton (2000), no se sigue una propuesta que atienda a revelar reglas sociales que regulan la producción y el acceso a los textos. Por el contrario, se legitiman rutinas escolares naturalizadas al afirmar que «las consignas se diseñaron considerando el género carta», lo cual se contradice con la afirmación de que «es el estudiante quien toma decisiones respecto de las prácticas de escritura que va a realizar» (Flotts et al., 2016a, pp.12-13). En tercer lugar, como fuente de esta evaluación de nivel primario, se recurre a un modelo de la enseñanza del lenguaje académico desde el nivel secundario —el de Snow y Uccelli (2009)— (cfr. Flotts et al., 2016a, p. 24) y no a desarrollos sobre la apropiación desde la infancia de diferentes géneros escritos. Por lo tanto, en *Aportes* se aprecian diferencias entre los modos en que la escritura es caracterizada y evaluada, lo que puede comprometer la relación entre la evidencia empírica obtenida y los sustentos teóricos.

4.2. Consignas

Las pruebas de escritura del TERCE tienen una consigna por grado. La consigna para tercer grado solicita escribir una carta sobre una situación en la que el estudiante evaluado habría participado con algunos amigos de su escuela, y la consigna de sexto grado pide escribir una carta formal en la que el estudiante fundamente la elección de una propuesta de edificación en su escuela. Estas consignas fueron planteadas con el propósito de generar «una situación comunicativa cercana, susceptible de ser real, con temáticas que no necesariamente involucran conocimientos especializados y con una intención, género y destinatario sencillos, claros, reconocibles» (Flotts et al., 2016a, p. 26).

Las consignas cumplen con varias de las características mencionadas en la sección *Referentes conceptuales*: incluyen el género (carta amical y de solicitud, respectivamente), el propósito (contar una experiencia y convencer de algo) y la audiencia (una amiga y una autoridad). No obstante, es posible problematizar la situación comunicativa, la familiaridad del género y la claridad de la formulación (Weigle, 2002).

En la consigna de tercer grado, puede cuestionarse la situación comunicativa en cuanto a la cercanía y el carácter no especializado pretendidos por el TERCE. En primer lugar, los estudiantes deben crear una situación ficticia para contar una experiencia sobre un concurso de música a una amiga ausente y, aunque escribir a un amigo puede resultar una situación cercana, hacerlo sobre una experiencia que no se ha tenido es poco o nada cotidiano o auténtico. En segundo lugar, debido al aumento de las comunicaciones mediante teléfonos celulares, la carta amical no necesariamente es un género familiar y puede comportar cierto nivel de especialización para algunos niños de tercer grado de primaria. Finalmente, en la consigna se indica: «Tu amiga te pide que le cuentes *cómo* estuvo la competencia y *cómo* le fue a tu grupo», y la palabra «cómo» puede dar lugar a una secuencia descriptiva en lugar de a la secuencia narrativa esperada; de hecho, el despliegue de «cómo es» es considerado base del prototipo textual descriptivo (Adam, 1992).

La consigna de sexto grado puede ser problematizada en el mismo sentido. Es posible que muchos estudiantes latinoamericanos se encuentren familiarizados con centros de padres y su capacidad para edificar plazas o cafeterías dentro de las escuelas, pero esto no se puede generalizar. Es relevante reflexionar sobre este aspecto, porque el conocimiento previo es fundamental para argumentar (Ferreti y Fan, 2017) y porque, con bastante frecuencia, se asume que las prácticas de escritura son independientes de los contextos cuando, en verdad, varían de acuerdo con normas y discursos culturales, y con posiciones dentro de relaciones de poder (Street, 2003). Además, en esta consigna, desde perspectivas lingüísticas como la de Snow y Uccelli (2009), tomada en *Aportes*, se advierte una complejidad que atentaría contra la sencillez buscada. En una misma cláusula, se observan un modificador del núcleo oracional, tres suboraciones y tres nominalizaciones que empacan procesos y agentes («*Para tomar la mejor decisión*, la presidenta del Centro de Padres ha pedido a cada curso *que entregue su opinión* en una carta, *en la que den buenas razones que apoyen su elección*»). Esto llama la atención, porque, según el *Reporte técnico*, uno de los criterios para la exclusión de ítems fue la falta de claridad, definida como «ítems que tienen en su redacción *uso de muchas subordinadas* o elementos de correferencia complejos para el grado» (Flotts et al., 2016b, p. 55).

4.3. Rúbricas

Las rúbricas empleadas en las pruebas de escritura del TERCE se estudiaron mediante un juicio de experto de los criterios, dominios y descriptores, así como de un análisis psicométrico de los instrumentos y sus ítems, el cual, como se indicó, complementa los análisis efectuados por el TERCE para las áreas de lectura, matemática y ciencias.

En *Aportes*, la exposición de las dimensiones e indicadores escogidos para la evaluación muestra una oscilación entre fundamentos teóricos. Por un lado, la justificación correspondiente a la dimensión de las «convenciones de legibilidad» se basa, como es esperable, en estudios que dan cuenta del desarrollo en los niños de «las habilidades de segmentación necesarias para el aprendizaje de la escritura» (Flotts *et al.*, 2016a, p. 23), como es el caso de las investigaciones de Emilia Ferreiro. Por otro lado, las dos dimensiones restantes («discursiva» y «textual») se fundamentan en el modelo de análisis del discurso de Martin y Rose (2003) y en las preocupaciones por la enseñanza y el aprendizaje del lenguaje académico planteadas por Snow y Uccelli (2009). Es decir, para estas dos dimensiones se desconocen fuentes que abordan el desarrollo de diferentes habilidades lingüísticas desde la niñez (*e.g.*, Beach, Newell y VanDerHeide, 2017). Quizás, sea esta la causa de que indicadores enunciados, para estas dos dimensiones en tercer grado de primaria, y puedan describir también escritos de estudiantes de nivel secundario; por ejemplo: «El texto está cohesionado: no faltan referentes ni se omiten palabras o nexos necesarios para la comprensión intra o interoracional. No hay contradicciones temáticas» (Flotts *et al.*, 2016a, p. 32). Puede suponerse que los manuales empleados por los codificadores proporcionan especificaciones para emplear las rúbricas, pero hasta ahora no han sido publicados.

En cuanto a los datos sobre las rúbricas, el *Manual para la elaboración de ítems* indica que «estas serán elaboradas en forma centralizada por el equipo de MIDE UC», sin precisar el proceso de validación. El apartado «Análisis de preguntas abiertas y escritura» del *Reporte técnico* anuncia el análisis de «los patrones de respuesta, la dificultad según TRI y las curvas características en cada ítem» (Flotts *et al.*, 2016b, p. 70), pero el apartado «Resultados del análisis: preguntas abiertas» (p. 90) no proporciona información psicométrica sobre las rúbricas y los ítems que las componen. De esta manera, a diferencia de lo que ocurre en las pruebas de lectura, matemática y ciencias, el *Reporte técnico* no brinda evidencias de validez basadas en la estructura interna de las rúbricas y tampoco, evidencias de confiabilidad. Dado que las puntuaciones de los estudiantes en los ítems de las rúbricas fueron empleadas para calcular

promedios nacionales y establecer comparaciones entre países, esta ausencia es llamativa.

Como parte del presente estudio, se realizó un análisis psicométrico de las rúbricas de tercer y sexto grado de primaria, y sus indicadores (ítems). El análisis de componentes principales de los residuos estandarizados proporciona evidencias en contra del supuesto de unidimensionalidad en ambas rúbricas. Según Linacre (2017), si el primer autovalor explica menos del 5% de la varianza o si su valor es menor que 3, no hay serias evidencias en contra del supuesto de unidimensionalidad. Los datos obtenidos indican que en tercer grado el primer autovalor explica el 11,8 % de la varianza y su valor es 1,56; y que en sexto grado el primer autovalor explica el 11,2 % de la varianza y su valor es 1,59.

En relación con la confiabilidad, el Alpha de Cronbach es 0,69 tanto en tercer grado como en sexto, y la separación de personas es 1,34 y 1,18 en tercer y sexto grado, respectivamente. Estos valores no alcanzan el mínimo aceptable de 0,70 (Barton y Peat, 2014) y 1,5 (Linacre, 2017), respectivamente.

Adicionalmente, se evaluaron el ajuste de los ítems al modelo de Rasch y la dificultad de los mismos. Según Linacre (2017), los valores de ajuste entre 0,5 y 1,5 son productivos para la medición; los valores entre 1,5 y 2, improductivos para la medición, pero no la distorsionan, y los valores superiores a 2 distorsionan la medición. Como puede apreciarse en la Tabla 1 y en la Tabla 2, casi todos los ítems ajustan al modelo.

En cuanto a la dificultad de cada ítem, el modelo de Rasch la relaciona con la habilidad de los evaluados (Linacre, 2017). Como se observa en la Tabla 1 y en la Tabla 2, la mayoría de los valores se aproxima a 0, lo cual indica muy poca variabilidad en la dificultad de los ítems en relación con la habilidad del conjunto de los evaluados de cada grado.

Tabla 1. Ajuste y dificultad de los ítems de la evaluación de escritura de tercer grado

Ítem	Ajuste próximo	Ajuste lejano	Dificultad
1	1,08	1,04	0,86
2	1,10	1,09	-0,06
3	0,83	0,84	-0,22
4	0,78	0,81	-0,94
5	0,82	0,88	0,05
6	0,74	0,82	-0,03
7	0,98	0,93	0,14
8	1,52	1,35	0,19

Fuente: elaboración propia a partir de datos del TERCE.

Tabla 2. Ajuste y dificultad de los ítems de la evaluación de escritura de sexto grado

Ítem	Ajuste próximo	Ajuste lejano	Dificultad
1	1,09	1,05	0,40
2	1,13	1,09	0,43
3	1,15	1,18	-0,02
4	0,78	0,69	-0,38
5	0,78	0,80	-0,53
6	0,74	0,80	0,12
7	0,80	0,86	-0,06
8	1,02	0,97	0,19
9	1,15	1,01	-0,14
10	1,28	1,15	-0,02

Fuente: elaboración propia a partir de datos del TERCE.

El análisis de la relación entre la habilidad de los estudiantes y la dificultad de los ítems a través de mapas de Wright muestra que las habilidades se ubican entre -3,5 y +3,9 en tercer grado y entre -3,5 y +3,8 en sexto, con distribuciones de mayor frecuencia en la parte positiva de la escala: en tercer grado, más del 70% de los evaluados se ubica por encima del valor medio y en sexto grado, sucede lo mismo con más del 88%. Al comparar estas habilidades con las dificultades de las preguntas (ninguna de las cuales sobrepasa el rango de -1 a +1), se observa falta de equiparación entre la habilidad y la dificultad. En ambos grados, y más marcadamente en sexto, los evaluados tienen más habilidad que la dificultad ofrecida por los ítems.

4.4. Proceso de codificación

En los documentos publicados, se halló una única referencia a la codificación de las pruebas de escritura del TERCE: la cantidad de escritos corregidos en el pilotaje de cada grado (Flotts *et al.*, 2016b, p. 116). Por tanto, se desconoce cuántos fueron los codificadores de la prueba definitiva, qué formación tenían y de qué modo fueron capacitados para codificar. Se ignora también con qué técnicas se controlaron eventuales sesgos y faltas de consistencia intercodificadores e intracodificadores. El reanálisis de ejemplos proporcionados en *Aportes* para ilustrar las diferencias entre niveles de desempeño refuerza el imperativo de consensuar y controlar la codificación. Por ejemplo, el siguiente texto de tercer grado es ubicado en el Nivel III por tener dos errores de Cohesión:

Querida Amiga:

En la competencia final después de que te cambiaste de ciudad ganamos el segundo lugar en el concurso de música el grupo estuvo muy feliz pero lo malo es que tu no estabas, pero nos dieron un trofeo de plata unas medallas y un diploma y hai acaba. ¡Adios!

Según *Aportes*, el primer error consiste en usar «pero» para establecer una relación que no es adversativa en «tú no estabas, pero dieron trofeo». No obstante, «las inferencias que determinan la interpretación de la conjunción, *pero* son de deducciones libres establecidas en función de muy diversas informaciones» (Real Academia Española, 2009, p. 2453). Así, cabe considerar que mediante «pero» el estudiante opone aceptablemente algo «malo» («no estabas») y algo bueno («nos dieron un trofeo»). El segundo error señalado es «al final de la carta (...) la repetición de la conjunción ‘y’» (Flotts *et al.*, 2016a, p. 66). Sin embargo, «Y ahí acaba» puede tomarse como un uso propio de la narración, como en «Y vivieron felices». Entonces, el texto se ubicaría en el Nivel IV y no III.

De manera similar, en la página 79, se ejemplifica el Nivel III de Registro con un texto que tendría una sola marca de oralidad e informalidad («*Bueno*, ya me despido»), pero el reanálisis evidencia cuatro errores más: «Señora Presidenta (...) *Querida* Presidenta (...) me dirijo a Usted para comunicarte» y «me despido con *un fuerte abrazo hasta luego*».

Más llamativo aún, el Nivel III de Puntuación, en sexto grado, es caracterizado por la presencia de «uno de los siguientes errores: interrumpir una oración con punto seguido; ausencia de coma en una enumeración; ausencia de punto final», y el ejemplo que se ofrece presenta tres enumeraciones no separadas por coma: 1) de motivos para querer una plaza («para jugar, para reunirse, etc.»); 2) de atributos de la plaza («con cancha de futbol, toboganes, amacas, sube y bajas, y mucha cosas más») y 3) de bienes de una cafetería («que tenga sillas, mesas, baños, comidas») (Flotts *et al.*, 2016a, p. 94). Estos y otros ejemplos no tratados aquí sugieren que los resultados en escritura podrían deberse también a fallas en la codificación.

5. DISCUSIÓN Y CONCLUSIONES

La confiabilidad tuvo mayor peso que la validez en las evaluaciones de la escritura durante el siglo XX (Lynne, 2004), en el cual se optó por la aproximación indirecta debido a la dificultad de obtener adecuados índices de consistencia entre los evaluadores. Las nuevas propuestas de evaluación buscan balancear los requerimientos de confiabilidad y validez (Bridgeman, 2016). No obstante,

persisten los desafíos (Zabaleta, Roldán y Centeleghe, 2016), por lo que, actualmente, la prueba de escritura del LLECE es una de las pocas evaluaciones internacionales que estudia la escritura de los estudiantes de nivel primario con aproximación directa. Este trabajo pretendió aportar a la iniciativa del LLECE al analizar el marco teórico, las consignas, las rúbricas y la codificación con técnicas cualitativas y cuantitativas usuales en la evaluación educativa.

El análisis del marco teórico evidenció una amplia revisión bibliográfica de teorías de la escritura, pero también, falta de articulación entre dicha revisión, los objetivos de las pruebas y sus características, así como cierta indiferenciación entre investigar y evaluar la escritura. Para fortalecer la relación entre la evidencia empírica obtenida y los sustentos teóricos planteados, futuros estudios podrían explicitar las características de la evaluación de la escritura a gran escala (Murphy y Blake Yancey, 2008), diferenciándola de las investigaciones aplicadas en psicología cognitiva (Bereiter y Scardamalia, 1992), nuevos estudios de literacidad (Barton y Hamilton, 2000), lingüística funcional (Martin y Rose, 2003) y otros enfoques.

El análisis de las consignas mostró que en su formulación se consideraron aspectos clave como el género, el destinatario y el propósito; no obstante, también reveló, en ambos grados, problemas relacionados con la cercanía contextual pretendida. Además, en la consigna de tercer grado, se identificó una palabra clave que remite a un tipo textual no previsto y, en sexto grado, una complejidad sintáctica que atenta contra las pautas de redacción fijadas por el propio TERCE. Siguiendo a Weigle (2002), debe prestarse atención a la descripción del contexto de escritura, y la claridad y precisión de las consignas en tanto constituyen dos de los aspectos más importantes del diseño de consignas de escritura.

En el *Reporte técnico*, se informa que se pilotearon dos consignas por grado, las cuales fueron revisadas y calificadas únicamente por el experto del área respectiva (Flotts *et al.*, 2016b), a diferencia de los ítems de las pruebas de lectura, ciencias y matemática, que recibieron cuatro revisiones más. Sin embargo, cada uno de los componentes de una consigna (género discursivo, tema, sintaxis) puede incidir por sí mismo en la escritura, y es fundamental pilotear y someter a análisis una serie amplia de consignas de escritura antes de decidir cuáles se incluirán en la evaluación (Bridgeman, Trapani y Bivens-Tatum, 2011).

Futuros estudios podrían poner a prueba en la aplicación piloto más cantidad de consignas para asegurar su calidad en la aplicación definitiva. Asimismo, para las consignas de las pruebas de escritura, debería acudirse a las mismas instancias de revisión que en otras áreas evaluadas, incluido un juicio

de expertos amplio. Así, siguiendo a Kane (2016), se obtendrían mayores evidencias de validez de contenido.

El análisis cualitativo de las rúbricas reveló que las dimensiones e indicadores escogidos para la evaluación se basan en estudios sobre el desarrollo de las habilidades de la escritura en la niñez en el caso de las convenciones de legibilidad, pero no en el caso de las dimensiones discursiva y textual. Futuras evaluaciones podrían considerar estudios sobre estas dos últimas dimensiones con una perspectiva de desarrollo; por ejemplo, Donovan y Smolkin, 2017; Christie y Derewianka, 2008 y van den Bergh, Rijlaarsdam y van Steendam, 2017. Del mismo modo, se observó que el *Reporte técnico* —a diferencia de las otras evaluaciones del TERCE— no presenta evidencias de validez basadas en la estructura interna de las rúbricas y tampoco, evidencias de confiabilidad, a pesar de que las puntuaciones de los estudiantes en los ítems fueron empleadas para calcular promedios nacionales y establecer comparaciones entre países. Con los datos disponibles, se realizó un análisis cuantitativo mediante el modelo Rasch de crédito parciales (Masters, 2016) que mostró evidencias en contra del supuesto de unidimensionalidad de acuerdo con Linacre (2017). Esto indica la necesidad de que en futuros estudios se revise la estructura dimensional de las rúbricas.

Además, se encontró que los índices de confiabilidad no alcanzan los mínimos aceptables sugeridos por Barton y Peat (2014) y Linacre (2017), lo cual significaría que las rúbricas no son lo suficientemente sensitivas para distinguir el desempeño de estudiantes de alta y baja habilidad. Para incrementar la confiabilidad de las rúbricas, futuros estudios podrían recoger información sobre el desempeño en escritura a través de un número mayor de indicadores (INEE, 2013) que permitan captar distintos niveles de habilidad.

El proceso de codificación se analizó solo de modo cualitativo debido a que las bases de datos del TERCE carecen de información sobre los codificadores. La revisión reveló ausencia de información sobre el modo en que estos fueron seleccionados y capacitados, así como alertó sobre la calidad del proceso. Futuros estudios del LLECE podrían incluir información de este tipo, el cual brinda evidencias de la validez de las interpretaciones y la confiabilidad de las puntuaciones (Eckes, 2015).

La prueba de escritura del LLECE representa un logro regional incuestionable. Las sugerencias surgidas del análisis apuntan a fortalecer esta y otras pruebas de la escritura en Latinoamérica.

REFERENCIAS BIBLIOGRÁFICAS

- ACE (2015). Informe técnico Simce 2013. Santiago de Chile. Recuperado de [http://www.agenciaeducacion.cl/wp-content/uploads/2016/03/Informe TecnicoSimce_2013.pdf](http://www.agenciaeducacion.cl/wp-content/uploads/2016/03/Informe_TecnicoSimce_2013.pdf)
- Adam, J.-M. (1992). *Les textes: types et prototypes. Récit, description, argumentation, explication et dialogue*. Paris: Nathan.
- AERA, APA y NCME (2014). *Standards for Educational & Psychological Testing*. Washington, DC.
- Atorresi, A. (2005). Construcción y evaluación de consignas para evaluar la escritura como competencia para la vida. *Enunciación*, 10(1), 4-14.
- Atorresi, A. y cols. (2010). *Escritura: Un estudio de las habilidades de los estudiantes de América Latina y el Caribe*. Santiago: OREALC/Unesco. Recuperado de <http://unesdoc.unesco.org/images/0019/001919/191925s.pdf>
- Bachman, L. (2002). *Some reflections on task-based language performance assessment*. *Language Testing*, 19(4), 453-476. <https://doi.org/10.1191/0265532202lt240oa>
- Barton, B. y Peat, J. (2014). *Medical Statistics: A Guide to SPSS, Data analysis and Critical Appraisal*. New Delhi: Wiley Blackwell.
- Barton, D. y Hamilton, M. (2000). Literacy practices. En D. Barton, M. Hamilton y R. Ivanič (eds.), *Situated Literacies. Reading and Writing in Context* (pp. 7-15). Londres: Routledge.
- Bazerman, C. (2017). What do sociocultural studies of writing tell us about learning to write? En C. MacArthur, S. Graham y J. Fitzgerald (eds.), *Handbook of Writing Research* (pp. 11-23). Nueva York: The Guilford Press.
- Beach, R; Newell, G.E. y VanDerHeide, J. (2017). A sociocultural perspective on writing development: Toward an agenda for classroom research on students' use of social practices. En C. MacArthur, S. Graham y J. Fitzgerald (eds.), *Handbook of Writing Research* (pp. 88-101). Nye York: The Guilford Press.
- Behizadeh, N. (2014). Mitigating the dangers of a single story: Creating large-scale writing assessments aligned with sociocultural theory. *Educational Researcher*, 43(3), 125-136. <https://doi.org/10.3102/0013189X14529604>
- Bejar, I. I. (2017). A historical survey of research regarding constructed-response formats. En R. E. Bennett y M. von Davier (eds.), *Advancing Human Assessment* (pp. 565-633). Nueva York: Springer Science. https://doi.org/10.1007/978-3-319-58689-2_18
- Bereiter, C. y Scardamalia, M. (1992). *Dos modelos explicativos de la composición escrita. Infancia y Aprendizaje*, (58), 43-64. <https://doi.org/10.1080/02103702.1992.10822332>

- Bond, T. G. y Fox, C. M. (2007). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*. Mahwah, NJ: LEA. <https://doi.org/10.4324/9781410614575>
- Bridgeman, B. (2016). *Can a two-question test be reliable and valid for predicting academic outcomes? Educational Measurement: Issues and Practice*, 35(4), 21-24. <https://doi.org/10.1111/emip.12130>
- Bridgeman, B.; Trapani, C. y Bivens-Tatum, J. (2011). Comparability of essay question variants. *Assessing Writing*, 16(4), 237-255. <https://doi.org/10.1016/j.asw.2011.06.002>
- Burga, A. (2006). La unidimensionalidad de un instrumento de medición: perspectiva factorial. *Revista de Psicología*, 24(1), 53-80. Recuperado de <http://revistas.pucp.edu.pe/index.php/psicologia/article/view/642>
- Calfee, R. y Miller, R. (2007). Best practices in writing assessment. En S. Graham, C. MacArthur y J. Fitzgerald (eds.), *Handbook of Writing Research* (pp. 265-280). Nueva York: The Guilford Press.
- Chan, S., Inoue, C. y Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, 26, 20-37. <https://doi.org/10.1016/j.asw.2015.07.004>
- Christie, F. y Derewianka, B. (2008). *School Discourse. Learning to Write across the Years of Schooling*. Londres, Nueva York: Continuum.
- Donovan, C. y Smolkin, L. (2017). Children's understanding of genre and writing development. En C. MacArthur, S. Graham, S. y J. Fitzgerald (eds.), *Handbook of Writing Research* (pp. 131-143). Nueva York: The Guilford Press.
- Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement*. Nueva York: Peter Lang Edition.
- Ferretti, R. y Fan, Y. (2017). Argumentative writing. En C. MacArthur, S. Graham y J. Fitzgerald (eds.), *Handbook of Writing Research* (pp. 301-315). Nueva York: The Guilford Press.
- Flotts, M. P., Manzi, J., Abarzúa, A., Cayumán, C., Treviño, E., Meyer, A., Jiménez Rodríguez, J. (2016b). *Tercer Estudio Regional Comparativo y Explicativo (TERCE). Reporte técnico*. Santiago: OREALC/Unesco. Recuperado de <http://www.unesco.org/new/es/santiago/education/education-assessment-llece/terce/>
- Flotts, M. P., Manzi, J., Jiménez, D., Abarzúa, A., Cayuman, C. y García, M.J. (2015). Los aprendizajes en escritura. En *Tercer Estudio Regional Comparativo y Explicativo (TERCE). Informe de resultados. Logros de aprendizaje* (pp. 95-145). Santiago: OREALC/Unesco. Recuperado de <http://www.unesco.org/new/es/santiago/education/education-assessment-llece/terce/>

- Flotts, M. P., Manzi, J., Lobato, P., Durán, M. I., Díaz, M. P. y Abarzúa, A. (2016a). *Aportes para la enseñanza de la escritura*. Santiago: OREALC/Unesco. Recuperado de <http://www.unesco.org/new/es/santiago/education/education-assessment-llece/terce/>
- Greenberg, K. (1992). Validity and reliability issues in the direct assessment of writing. *Writing Program Administration* (16), 7-22. Recuperado de <https://eric.ed.gov/?id=EJ452781>
- Huot, B. (2002). *(Re) articulating Assessment: Writing Assessment for Teaching and Learning*. Logan: Utah State University Press.
- INEE (2013). *El aprendizaje de la expresión escrita en la educación básica en México*. México DF: INEE. Recuperado de <http://www.inee.edu.mx/index.php/72-publicaciones/resultados-de-aprendizaje-capitulos/567-el-aprendizaje-de-la-expresion-escrita-en-la-educacion-basica-en-mexico-sexto-de-primaria-y-tercero-de-secundaria>
- Johnson, R., Penny, J. y Gordon, B. (2009). *Assessing Performance: Designing, Scoring, and Validating Performance Tasks*. Nueva York: The Guilford Press.
- Jonsson, A., Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144. <https://doi.org/10.1016/j.edurev.2007.05.002>
- Kane, M. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. En S. Lane, M. Raymond y T. Haladyna (eds.), *Handbook of Test Development* (pp. 64-80). Nueva York: Routledge.
- Li, J. (2014). Examining genre effects on test takers' summary writing performance. *Assessing Writing*, 22, 75-90. <https://doi.org/10.1016/j.asw.2014.08.003>
- Linacre, J. M. (2017). Winsteps® (Version 4.0.0) [Computer Software]. Beaverton: Oregon. Recuperado de <http://www.winsteps.com/>
- Lynne, P. (2004). *Coming to Terms: A Theory of Writing Assessment*. Logan: Utah State University Press. <https://doi.org/10.2307/j.ctt46nqsz>
- Márquez, M. S., Iparraguirre, M. S. y Bengtsson, A. (2015). Un recorrido por el aprendizaje de la escritura desde una perspectiva evolutivo-educativa y semiótica. *Interdisciplinaria* 32(1), 151-168. <https://doi.org/10.16888/interd.2015.32.1.8>
- Martin, J. R. y Rose, D. (2003). *Interpreting social discourse. Working with Discourse: Meaning Beyond the Clause*. Londres: Continuum.
- Masters, G. N. (2016). Partial credit model. En W. van der Linden (ed.), *Handbook of Item Response Theory* (pp. 137-154). Boca Raton: Chapman & Hall/CRC.
- Mendoza, A. (2018). El uso de Many-Facet Rasch Measurement para examinar la calidad del proceso de corrección de pruebas de desempeño. *Revista*

- mexicana de investigación educativa*, 23(77), 597-625. Recuperado de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-66662018000200597&lng=es&nrm=iso
- Minedu (2016). *¿Qué logran nuestros estudiantes en Escritura?* Recuperado de <http://umc.minedu.gob.pe/evaluacion-censal-de-estudiantes-ece-2015/>
- Murillo, F. J. y Martínez-Garrido, C. (2016). La educación en América Latina y el Caribe. Aportes del TERCE y sus reanálisis. *REICE*, 14(4), 5-8. Recuperado de <https://revistas.uam.es/index.php/reice/article/view/6577>
- Murphy, S. y Blake Yancey, K. (2008). Construct and consequence: Validity in writing assessment. En C. Bazerman (ed.), (2009). *Handbook of Research on Writing: History, Society, School, Individual, Text* (pp. 448-474). Nueva York-Londres: Taylor & Francis Group.
- Real Academia Española (2009). *Nueva gramática de la lengua española*. Madrid: Espasa.
- Ruiz-Funes, M. (2015). Exploring the potential of second/foreign language writing for language learning: The effects of task factors and learner variables. *Journal of Second Language Writing*, 28, 1-19. <https://doi.org/10.1016/j.jslw.2015.02.001>
- Snow, C. y Uccelli, P. (2009). The challenge of academic language. En D. Olson y N. Torrance (eds.), *The Cambridge Handbook of Literacy* (pp. 112-133). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511609664.008>
- Street, B. (2003). The implications of “New Literacy Studies” for literacy education. En S. Goodman et al. (eds.), *Language Literacy and Education: A Reader* (pp.77-88). Stoke on Trent: Trentham Books.
- Tolchinsky, L. (2017). From text to language and back: The emergence of written language. En C. MacArthur, S. Graham y J. Fitzgerald (eds.), *Handbook of Writing Research* (pp.144-159). Nueva York-Londres: Guilford.
- Unesco y MIDE U.C. (2011). *Tercer Estudio Regional Comparativo y Explicativo (TERCE)*. Manual para la elaboración de ítems. (s/d).
- Van den Bergh, H., Rijlaarsdam, G. y van Steendam, E. (2017). Writing process pheory: A functional dynamic approach. En C. MacArthur, S. Graham y J. Fitzgerald (eds.), *Handbook of Writing Research* (pp. 57-71). Nueva York-Londres: Guilford Press.
- Villegas, F. (2016). *¿Qué logran los estudiantes peruanos en escritura?* En V. Villarán (Presidencia), V Seminario Nacional de Investigación Educativa. Ayacucho, Perú.
- Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511732997>

- Wiggins, G. (1998). *Educative Assessment. Designing Assessments to Inform and Improve Student Performance*. San Francisco: Jossey-Bass Publishers.
- Wind, S. A. y Peterson, M. E. (2017). *A systematic review of methods for evaluating rating quality in language assessment*. *Language Testing*, 35(2), 161-192. <https://doi.org/10.1177/0265532216686999>
- Zabaleta, V., Roldán, L. Á. y Centeleghe, M. E. (2016). *La evaluación de la escritura a la finalización de la escolaridad primaria según las pruebas TERCE y ONE*. *REICE*, 14(4). <https://doi.org/10.15366/reice2016.14.4.008>