CONFIABILIDAD Y VALIDEZ EN EVALUACIONES DE APRENDIZAJE DE ESTUDIANTES

ANDREA ABARZÚA – JOHANA CONTRERAS

Centro UC
Medición - MIDE

CONTENIDOS

I. Contexto

II. Acciones realizadas

III. Resultados

IV. Próximos desafíos

I. CONTEXTO

LAS PRUEBAS SEPA

- El Sistema de Evaluación de Progreso del Aprendizaje (SEPA) ha sido desarrollado desde el 2007 por el Centro de Medición MIDE UC de la Pontificia Universidad Católica de Chile.
- Consiste en un conjunto de pruebas estandarizadas, orientados a medir el nivel de aprendizaje de estudiantes respecto del currículum vigente en los sectores de Lenguaje y Matemática, desde 1° básico a III medio. El formato escogido por SEPA son las preguntas cerradas de selección múltiple, en papel y lápiz.
- El programa, en su conjunto, contempla además, acciones de asesoría y acompañamiento de los usuarios para la gestión y uso de los resultados de aprendizaje.
- El objetivo de SEPA es entregar información confiable, oportuna y útil para la toma de decisiones pedagógicas y de gestión.

LAS PRUEBAS SEPA

- El año 2007 se aplicaron aprox. 14.800 pruebas a 7.400 estudiantes pertenecientes a 37 establecimientos educacionales. El año 2017 se aplicaron un total de 104.500 pruebas, evaluando a aprox. 50.900 estudiantes y comprendiendo 234 establecimientos educacionales (51% municipales, 28% particulares subvencionados y 21% particulares pagados, de 13 regiones del país).

CONFIABILIDAD Y VALIDEZ

- Validez es el grado en el cual la evidencia y la teoría apoyan las interpretaciones de los puntajes de una prueba para los usos propuestos de las evaluaciones (AERA, APA & NCME, 2014).

- La validez es una propiedad en evolución, y la validación un proceso continuo (Martínez-Rizo, 2016).

- La confiabilidad (opuesta a error) es una propiedad requisito para la validez de una medición (AERA, APA & NCME, 2014).

II. ACCIONES REALIZADAS

Agenda de validación

Agenda de validación rutinaria

- Argumentos basados en el contenido
- Validez de estructura interna
- Verificaciones de Confiabilidad
- Verificaciones de Ecuanimidad

Agenda de estudios ocasionales

- Relación con otras variables
- Usos como evidencia de validez

III. RESULTADOS

AGENDA DE VALIDACIÓN RUTINARIA

ARGUMENTOS BASADOS EN EL CONTENIDO DE LAS PRUEBAS SEPA

- Tablas de especificaciones diseñadas y evaluadas por expertos disciplinarios orientadas a asegurar un *muestreo cuidadoso del contenido y las habilidades* (Koretz, 2010).
- Revisiones sucesivas de los ítems por especialistas en evaluación y expertos curriculares orientadas a *minimizar la sub-representación del constructo o la introducción de varianza irrelevante* (AERA, APA & NCME, 2014).
- Proceso iterativo de construcción: selección y capacitación de constructores de ítems; asesoría durante la elaboración y diseño de preguntas; revisión directa por especialistas y una revisión por una comisión técnica.
- Los criterios de estas revisiones son explícitos: rigurosidad conceptual, el alineamiento curricular y el cumplimiento de características de calidad en la formulación de los ítems. Los ítems, al final del proceso, son sometidos a la revisión de que realizan un dictamen sobre la continuidad de los ítems, en base a tres posibilidades: “aprobación”, “aprobación con modificaciones” o “rechazo”.

VALIDEZ DE ESTRUCTURA INTERNA DEL TEST

Lenguaje

Nivel	ChiSq PValue	CFI	TLI	RMSEA
1	0.00	0.96	0.95	0.03
2	0.00	0.96	0.96	0.03
3	0.00	0.96	0.96	0.03
4	0.00	0.98	0.98	0.02
5	0.00	0.96	0.96	0.03
6	0.00	0.97	0.97	0.03
7	0.00	0.97	0.97	0.02
8	0.00	0.97	0.97	0.02
9	0.00	0.96	0.96	0.02
10	0.00	0.97	0.97	0.02
11	0.00	0.92	0.92	0.03

Matemática

Nivel	ChiSq PValue	CFI	TLI	RMSEA
1	0.00	0.98	0.98	0.02
2	0.00	0.98	0.98	0.02
3	0.00	0.97	0.97	0.02
4	0.00	0.98	0.98	0.03
5	0.00	0.97	0.97	0.03
6	0.00	0.98	0.98	0.02
7	0.00	0.98	0.98	0.02
8	0.00	0.97	0.97	0.02
9	0.00	0.96	0.96	0.03
10	0.00	0.96	0.96	0.03
11	0.00	0.95	0.95	0.02

VERIFICACIONES DE CONFIABILIDAD

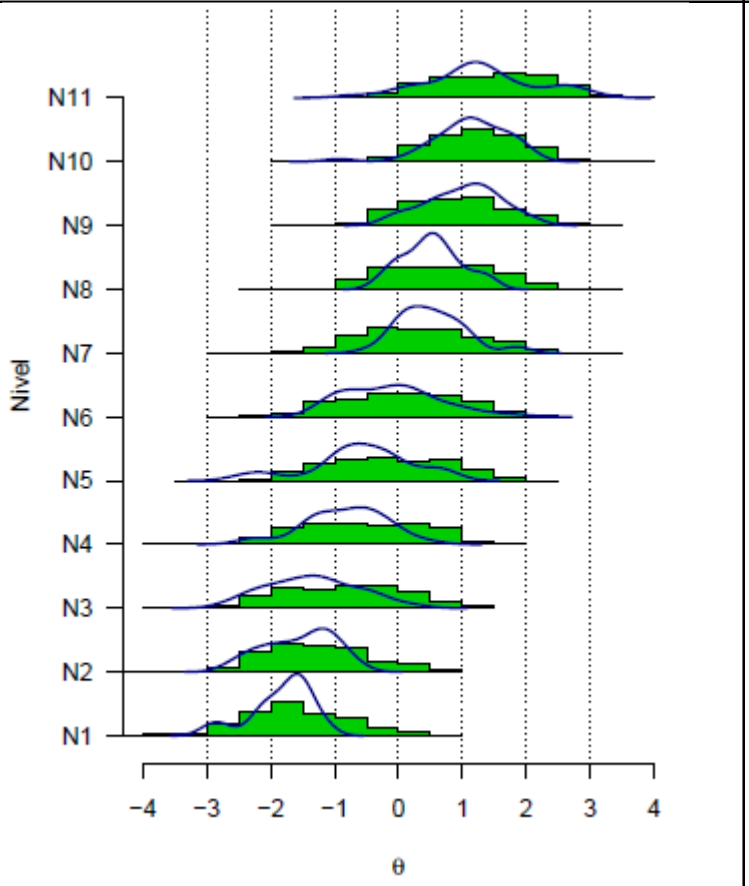
Lenguaje	2016	2017
1° Básico	0.82	0.82
2° Básico	0.85	0.84
3° Básico	0.88	0.89
4° Básico	0.90	0.90
5° Básico	0.88	0.89
6° Básico	0.89	0.88
7° Básico	0.89	0.90
8° Básico	0.88	0.90
I° Medio	0.85	0.88
II° Medio	0.87	0.86
III° Medio	0.86	0.90

Matemática	2016	2017
1° Básico	0.82	0.85
2° Básico	0.83	0.83
3° Básico	0.86	0.86
4° Básico	0.91	0.90
5° Básico	0.91	0.90
6° Básico	0.89	0.89
7° Básico	0.90	0.90
8° Básico	0.90	0.91
I° Medio	0.90	0.89
II° Medio	0.92	0.92
III° Medio	0.89	0.93

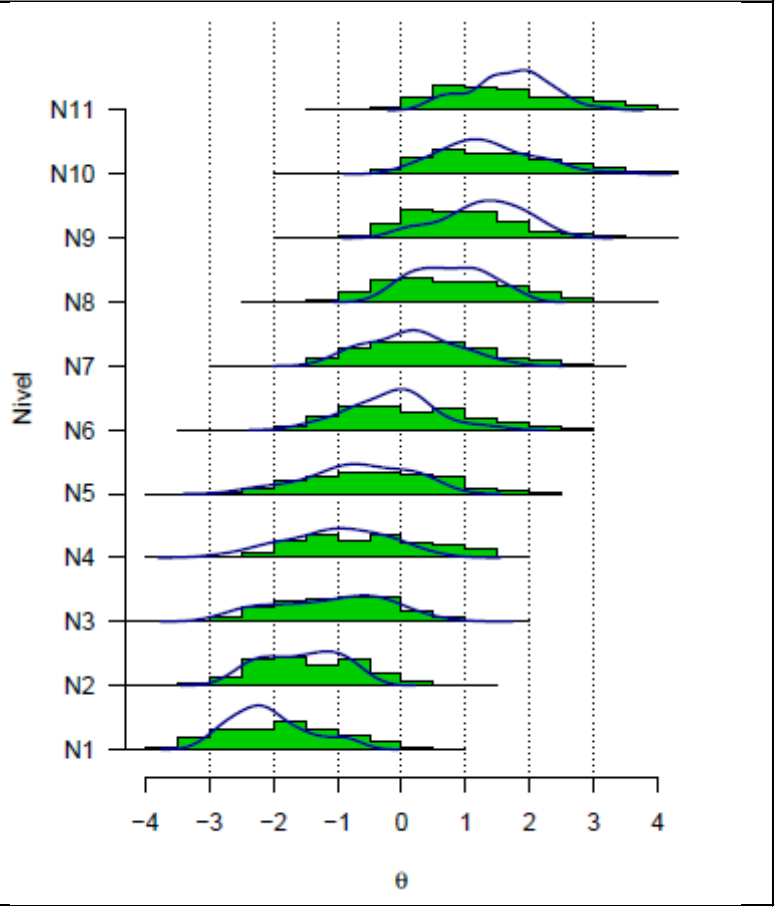
Lenguaje		Matemática	
2016	2017	2016	2017
0.93	0.92	0.95	0.95

VERIFICACIONES DE CONFIABILIDAD

Escala vertical Lenguaje



Escala vertical Matemática



VERIFICACIONES DE ECUANIMIDAD

se utilizan los criterios establecidos por ETS para clasificar ítems según el grado en que muestren evidencia de funcionamiento diferencial (Zwick, 2012).

Categoría	Comportamiento diferencial	Criterio
A	Despreciable	$P(X^2_{MH}) > 0.05$ ó $\lvert\alpha_{MH}\rvert \leq 1$
B	Ligero o Moderado	$P(X^2_{MH}) \leq 0.05$ y $1 < \lvert\alpha_{MH}\rvert < 1.5$
C	Moderado a Grande	$P(X^2_{MH}) \leq 0.05$ y $\lvert\alpha_{MH}\rvert \geq 1.5$

Lenguaje

Categoría	Frecuencia	Porcentaje
A+	233	50.7%
A–	218	47.4%
B+	1	0.2%
B–	5	1.1%
C+	1	0.2%
C–	2	0.4%

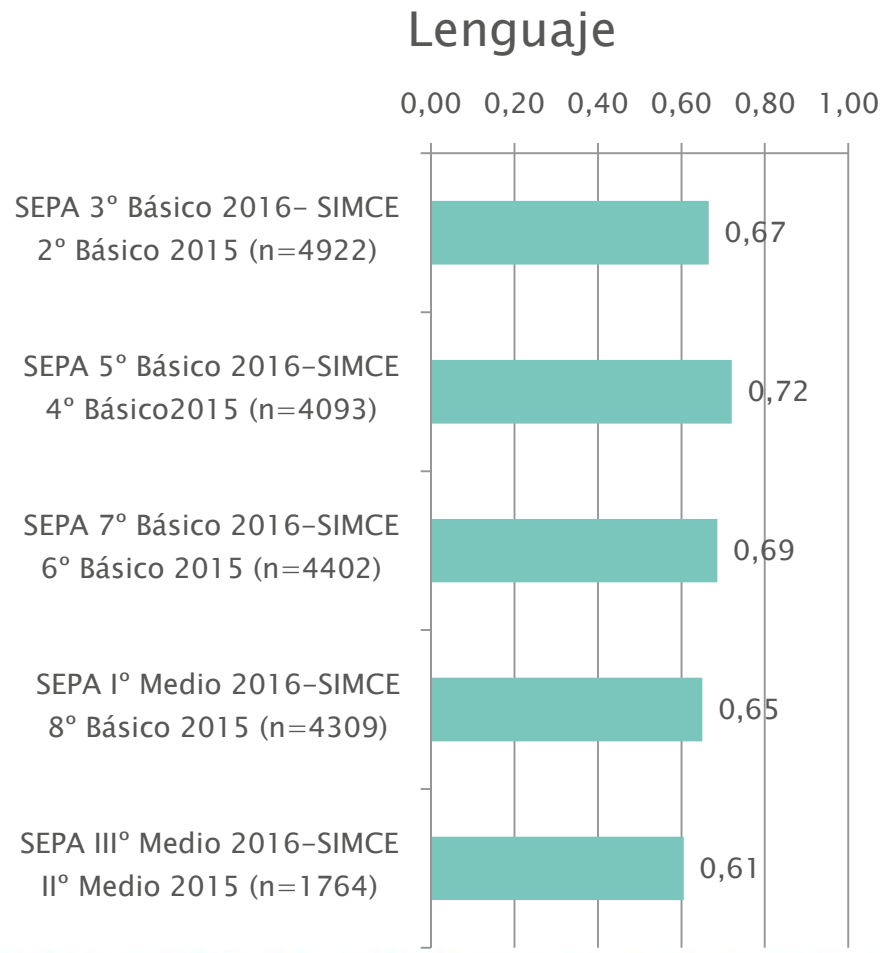
Matemática

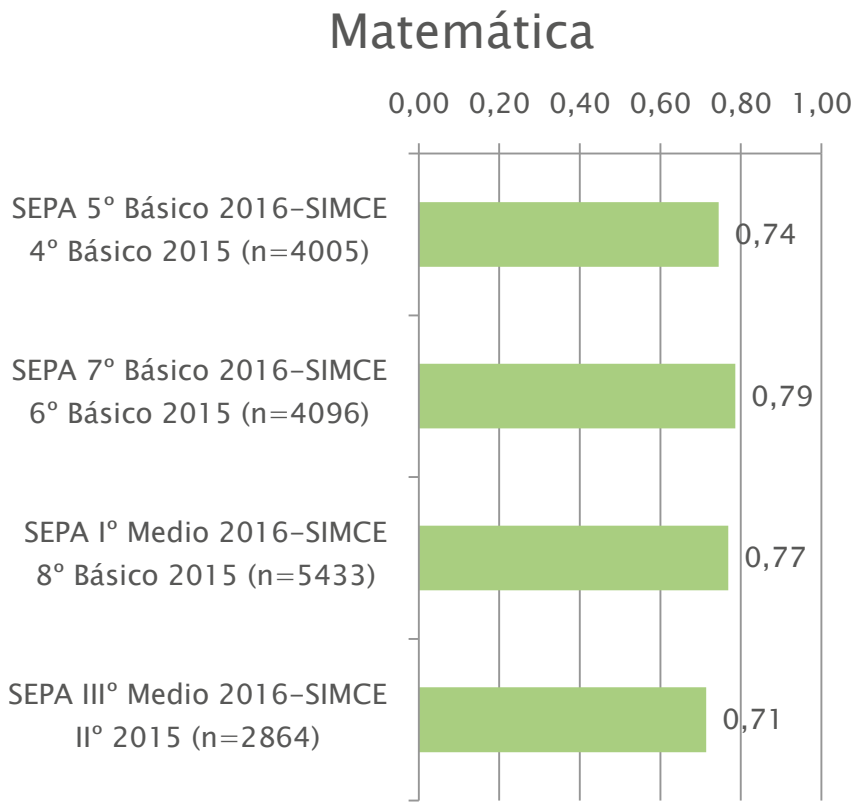
Categoría	Frecuencia	Porcentaje
A+	224	48.7%
A–	214	46.5%
B+	8	1.7%
B–	11	2.4%
C–	3	0.7%

GF=Niñas (+)

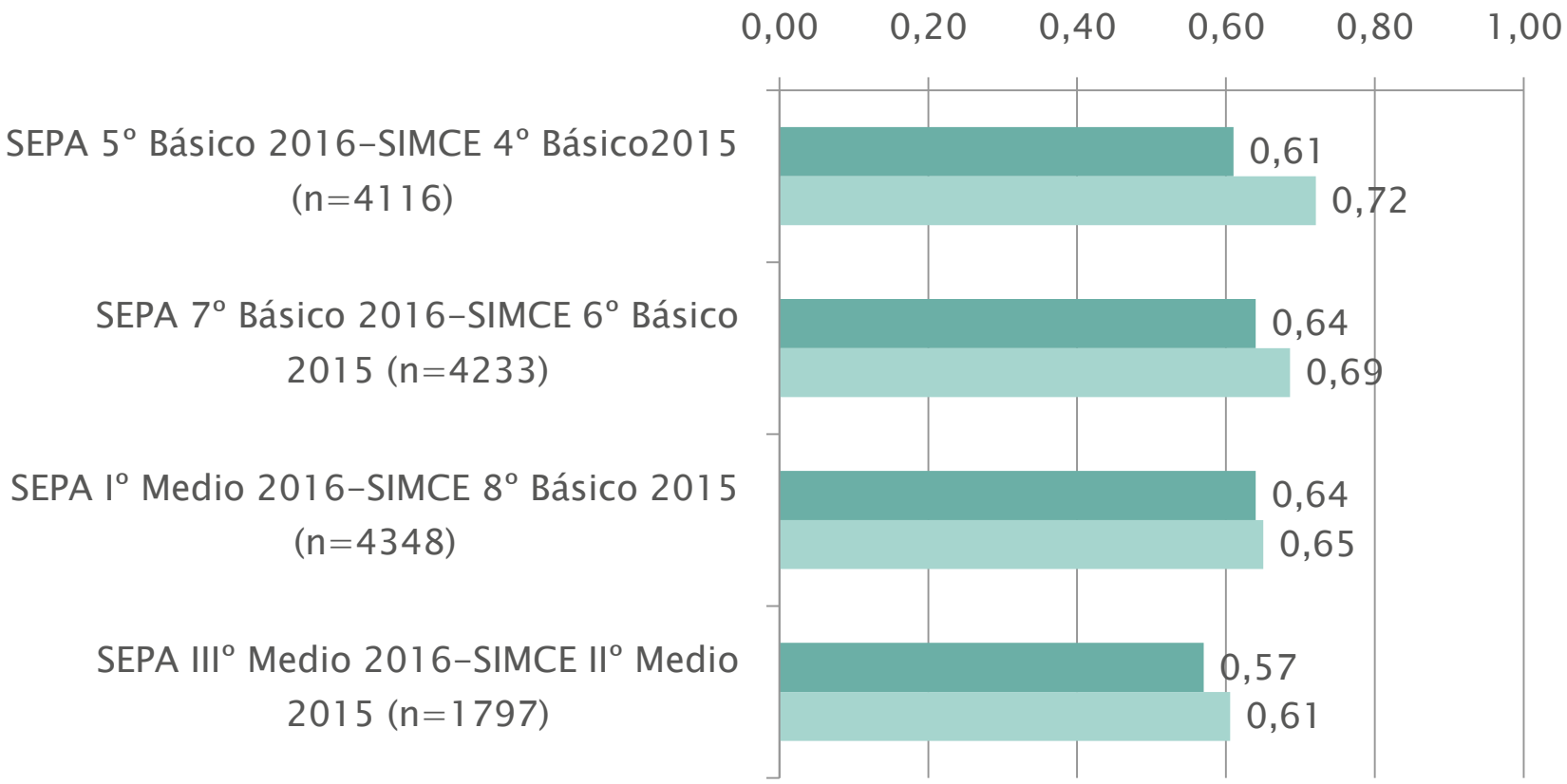
AGENDA DE VALIDACIÓN ESPACIADA

EVIDENCIA SOBRE LA RELACIÓN CON OTRAS VARIABLES: CORRELACIÓN CON SIMCE

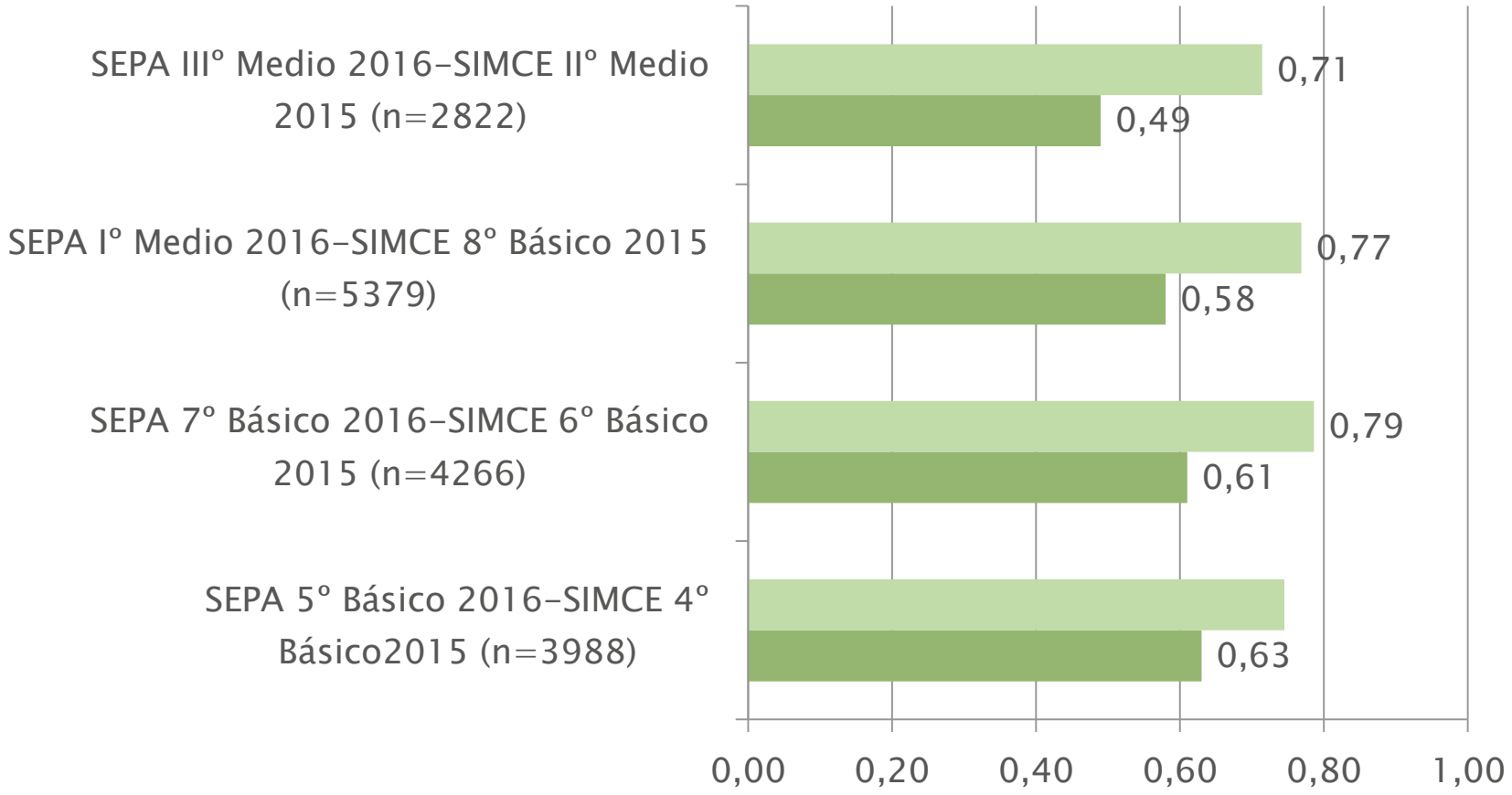




CORRELACIÓN PRUEBAS DE LENGUAJE SEPA- MATEMÁTICA SIMCE, RESPECTO DE CORRELACIÓN LENGUAJE SEPA-SIMCE



CORRELACIÓN PRUEBAS DE MATEMÁTICA SEPA- LENGUAJE SIMCE, RESPECTO DE CORRELACIÓN MATEMÁTICA SEPA-SIMCE



ESTUDIOS DE USOS COMO EVIDENCIA DE VALIDEZ CONSECUENCIAL

1. AGENDA 2017-2018: Fase de obtención de evidencia extraída de fuentes directas, respecto de los usos y consecuencias del programa. De esta manera, ha sido posible analizar el grado de concordancia entre los resultados esperados desde sus diseñadores y los usos que efectivamente se da a la información que se produce.
2. AGENDA 2017-2018: Incluye dos procesos: a) Difusión mediante presentaciones y publicaciones; b) Plan de trabajo interno a SEPA (retroalimentación y desafíos de mejora).

IV. PRÓXIMOS DESAFÍOS

Agenda rutinaria

- Estudios de validez de procesos de respuesta
- Estudios de ecuanimidad en fase piloto (otras categorías)

Agenda espaciada

- Agenda 2018-2019 estudios de usos (otros actores)
- Estudio de correlación con ERCE

REFERENCIAS

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*(2-3), 162-172.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 5.

Brewer, C., Knoeppel, R. C., & Lindle, J. C. (2015). Consequential validity of accountability policy: Public understanding of assessments. *Educational Policy*, *29*(5), 711-745.

Briggs, D. (2017). Learning Theory and Psychometrics: Room for Growth, Assessment in Education: Principles. *Policy & Practice*. *24*(3), 351-358. Doi: 10.1080/0969594X.2017.1336987

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.

Camilli, G. (2006). *Test fairness*. In R.Brennan (Ed.), Educational measurement. Westport , CT : American Council on Education and Praeger Publishing.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297-334.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and psychological measurement*, *64*(3), 391-418.

Cupani, M., & Zalazar-Jaime, M. F. (2014). Complex Traits and Academic Performance: Contribution of Personality Traits, Self-Efficacy, and Interests. *Revista Colombiana de Psicología*, *23*(1), 57-71.

REFERENCIAS

Haladyna, T. (2016). Item Analysis for Selected-reponse Test Items. En S. Lane, M. Raymond & T. Haladyna (Eds.). Handbook of Test Developpement. (2nd ed), pp. 392-409. New York: Routledge.

Hein, A. & Taut, S. (2010). El uso de información evaluativa externa con fines formativos: el caso de establecimientos educacionales chilenos participantes de SEPA. Revista Iberoamericana de Evaluación Educativa, 3(2), 160-181.

Heritage, M. (2010). Formative Assessment: Making it Happen in the Classroom. California: Corwin.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural equation modeling: a multidisciplinary journal, 6(1), 1-55.

Kane, M. (2016). Validation Strategies: Delineating and Validating Proposed Interpretations. En S. Lane, M. Raymond & T. Haladyna (Eds.). Handbook of Test Developpement. (2nd ed), pp. 64-80. New York: Routledge.

Manzi, J., Bogolasky, F., Gutiérrez, G., Grau, V. & Volante, P. (2014). Análisis sobre valoraciones, comprensión y uso del SIMCE por parte de directores escolares de establecimientos subvencionados. Santiuago: FONIDE.

Martínez-Rizo, F. (2016). Impacto de las pruebas en gran escala en contextos de débil tradición técnica: Experiencia de México y el Grupo Iberoamericano de PISA. RELIEVE, 22(1). DOI: http://dx.doi.org/10.7203/relieve.22.1.8244.

Morris, A. (2011). Student Standardised Testing: Current Practices in OECD Countries and a Literature Review. OECD Education Working Papers. N° 65, OECD Publishing. http://dx.doi.org/10.1787/5kg3rp9qbnr6-en

REFERENCIAS

Nkwake, A. (2015). Credibility, Validity, and Assumptions in Program Evaluation Methodology. Suiza: Springer.

Ravela, P., Picaroni, B. & Loureiro, G. (2017). ¿Cómo mejorar la evaluación en el aula? Montevideo: Grupo Magro.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. Educational and Psychological Measurement, 74(1), 31-57.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. Methods of psychological research online, 8(2), 23-74.

Taut, S., Santelices, M. V., Araya, C., & Manzi, J. (2011). Perceived effects and uses of the national teacher evaluation system in Chilean elementary schools. Studies in Educational Evaluation, 37(4), 218-229.

Taut, S., Santelices, V., Araya, C., & Manzi, J. (2010). Theory underlying a national teacher evaluation program. Evaluation and Program Planning, 33(4), 477-486.

Wise, L. & Plake, B. (2016). Test Design and Development Following the Standards for Educational. En S. Lane, M. Raymond & T. Haladyna (Eds.). Handbook of Test Developpement. (2nd ed), pp. 19-39. New York: Routledge.

Wylie, E. C. (2017). Winsight™ Assessment System: Preliminary Theory of Action. ETS Research Report Series.

Yu, C. Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes (Vol. 30). Los Angeles: University of California, Los Angeles.

¡Gracias por su atención!

Centro UC
Medición - MIDE

Andrea Abarzúa
Psicóloga y Ms. en Psicología Educacional de la
P. Universidad Católica de Chile.
raabarzu@uc.cl

Johana Contreras
Psicóloga y PhD. en Sociología de la
Universidad de Bordeaux, Francia.
jtcontre@uc.cl