Diversidad Cultural y Validez en las Pruebas de Aprovechamiento Escolar: Retos Prácticos y Posibilidades Metodológicas

Guillermo Solano-Flores
Stanford University

Seminario, Evaluación y Diversidad: Justicia y Equidad en la Evaluación Educativa"
Instituto Nacional para la Evaluación de la Educación
Ciudad de México, 22 de septiembre de 2016

Definiciones

- *Cultura.* Conjunto de valores, experiencias, patrones de comunicación, formas de socialización y circunstancias históricas que comparten los individuos de un grupo social.*
- *Validez cultural.* Grado de efectividad con el que el proceso de evaluación considera las influencias socioculturales en la manera en que los estudiantes interpretan los ítems de una prueba y responden a esos ítems. Tales influencias incluyen: experiencias sociales diarias, patrones de comunicación, variedades de lenguaje, epistemologías, y condiciones socioeconómicas. **

* Consejo Técnico Especializado Ad-Hoc para la Elaboración de Criterios Técnicos de Validez Cultural de los Instrumentos de Evaluación Educativa (2015). *Promoción y evaluación de la validez cultural en las actividades evaluativas del INEE*. (Por Orden Alfabético de Apellido): Gigante, E. von Groll, B. Martinez-Casas, R., Sandoval-Cruz, F. y Solano-Flores, G. (2015). Instituto Nacional para la Evaluación de la Educación Ciudad de México, D.F., Enero 16.

** Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38(*5*)*, 553-573.

Limitaciones en las prácticas evaluativas

- Se piensa en diversidad cultural y lingüística al final del proceso de desarrollo de la prueba
- Se asume universalidad en el diseño de ítems
- Se invoca la existencia de técnicas estadísticas para analizar sesgo, sin que realmente se les emplee
- Se trata a la diversidad cultural como un asunto superficial, no técnico
- Se asignan pocos recursos y poco tiempo para examinar asuntos de cultura
- Se usan o promueven estereotipos sociales
- Se subestima la complejidad de los asuntos culturales y lingüísticos
- Se hace algo con respecto a cultura por conveniencia política, sin un compromiso profundo
- Se considera que las acciones para tratar cultura son opcionales
- Se operacionaliza a la cultura de manera inadecuada

Factores agravantes

- Pocos expertos en medición están interesados en conocer a fondo las bases teóricas de la relación entre cognición, cultura y lengua
- Pocos expertos en antropología y lingüística están interesados en conocer a fondo las bases teóricas y metodológicas de la medición
- Se establece una falsa dicotomía entre lo cualitativo y lo cuantitativo
- En la mayoría de los países, existen pocos expertos en medición
- Muchos países importan prácticas evaluativas de otros países, incluyendo aquellas que tienen que ver con diversidad cultural, sin cuestionar sus limitaciones

Buena ideas, dudosa implementación
Estándar 3.3: Inclusión de sub-grupos poblacionales

Comment: Test developers should include individuals from relevant subgroups of the intended testing population in pilot or field test samples used to evaluate item and test appropriateness for construct interpretations. The analyses that are carried out using pilot and field testing data should seek to detect aspects of test design, content, and format that might distort test score interpretations for the intended uses of the test scores for particular groups and individuals. Such analyses could employ a range of methodologies, including those appropriate for small sample sizes, such as expert judgment, focus groups, and cognitive labs. Both qualitative and quantitative sources of evidence are important in evaluating whether items are psychometrically sound and appropriate for all relevant subgroups.

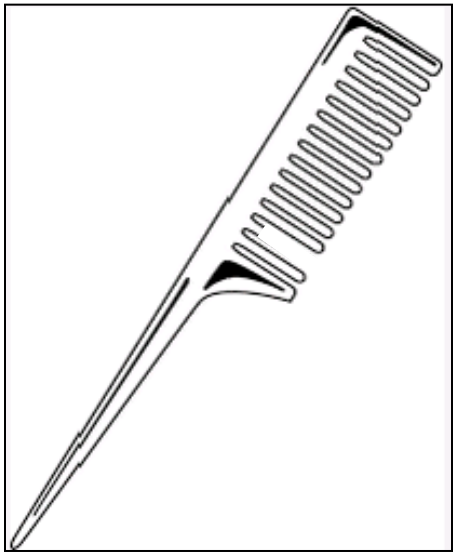
If sample sizes permit, it is often valuable to carry out separate analyses for relevant subgroups of the population. When it is not possible to include sufficient numbers in pilot and/or field test samples in order to do separate analyses, operational test results may be accumulated and used to conduct such analyses when sample sizes become large enough to support the analyses.

If pilot or field test results indicate that items or tests function differentially for individuals from, for example, relevant age, cultural, disability, gender, linguistic and/or racial/ethnic groups in the population of test takers, test developers should investigate aspects of test design, content, and format (including response formats) that might contribute to the differential performance of members of these groups and, if warranted, eliminate these aspects from future test development practices.

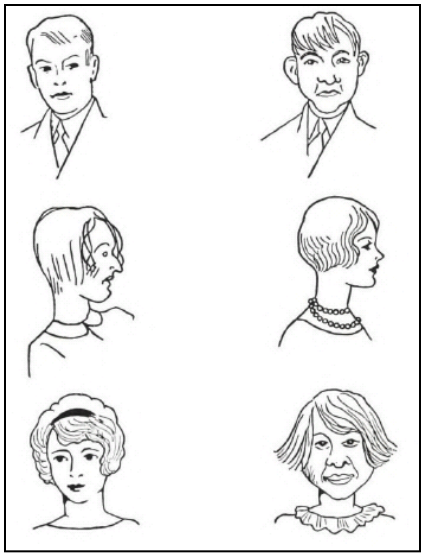
Expert and sensitivity reviews can serve to guard against construct-irrelevant language and images, including those that may offend some individuals or subgroups, and against construct-irrelevant context that may be more familiar to some than others. Test publishers often conduct sensitivity reviews of all test material to detect and remove sensitive material from tests (e.g., text, graphics, and other visual representations within the test that could be seen as offensive to some groups and possibly affect the scores of individuals from these groups). Such reviews should be conducted before a test becomes operational.

American Educational Research Association, American Psychological Association, and National Council for Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Ch. 3, Fairnes in Testing, p. 64. Washington, DC: Author.

Un ejemplo histórico sobre sesgo cultural en pruebas



*¿Qué le falta?**



*¿En cada par de caras, cuál es más bonita?***

*Basado en y **citado por Jensen, A. R. (1960). Bias in mental testing. New York: The Free Press.

Cultura y contexto en ítems

Item:
Juan va a cenar con sus papás. Su papá deja $17.00 de propina. ¿Cuál es el total de la cuenta de la cena, suponiendo que se agrega el 10 % de propina a la cuenta total?

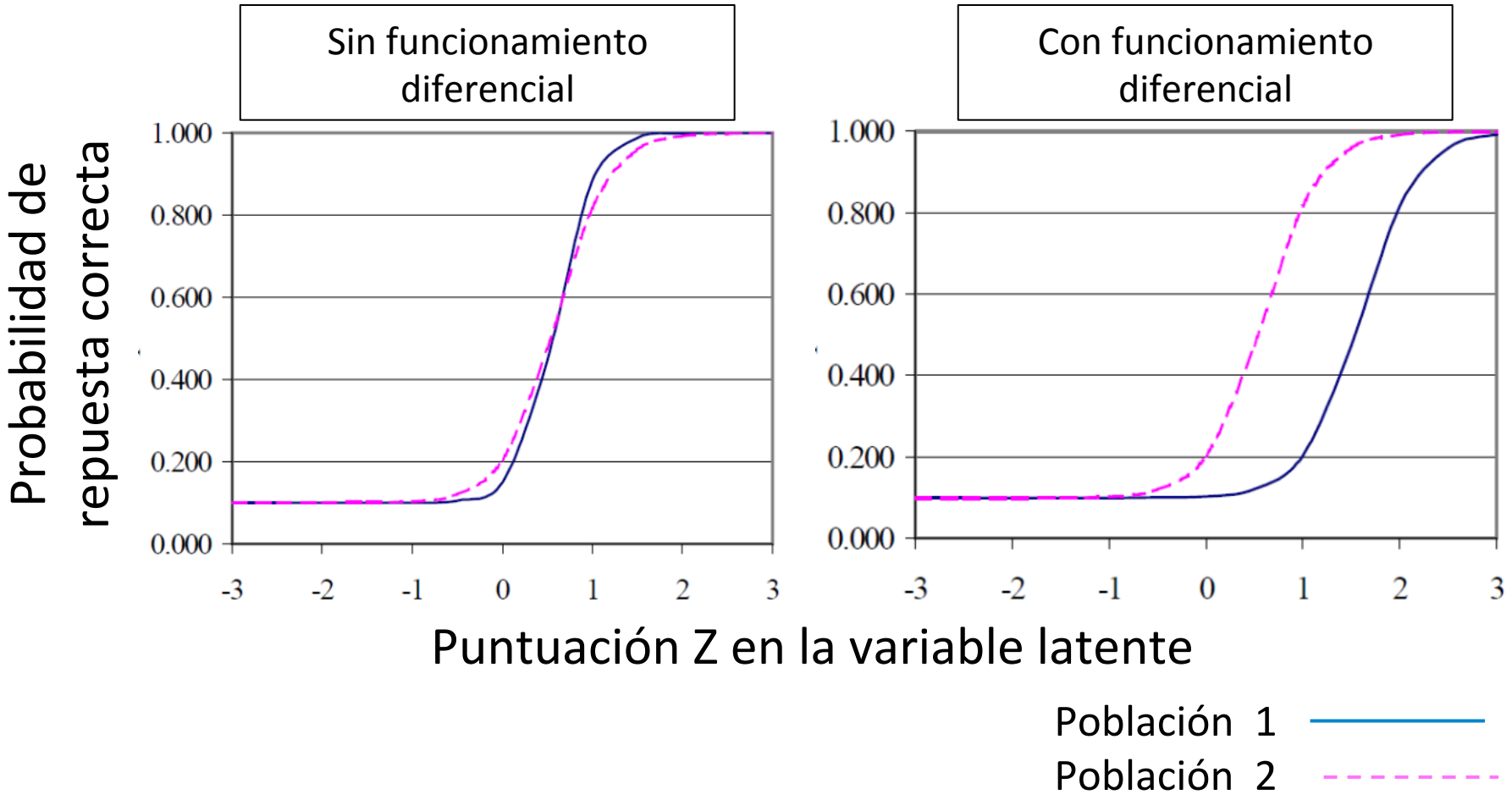
Reporte técnico:

Un panel de expertos revisó los ítems para asegurarse de que no resultaran difíciles de entender para alumnos de ciertos grupos culturales o lingüísticos.

Nos aseguramos de usar un lenguaje estándar en la redacción de los ítems

Empleamos los principios del diseño universal

Enfoques convencionales en el tratamiento de cultura: Ejemplo de DIF en dos ítems

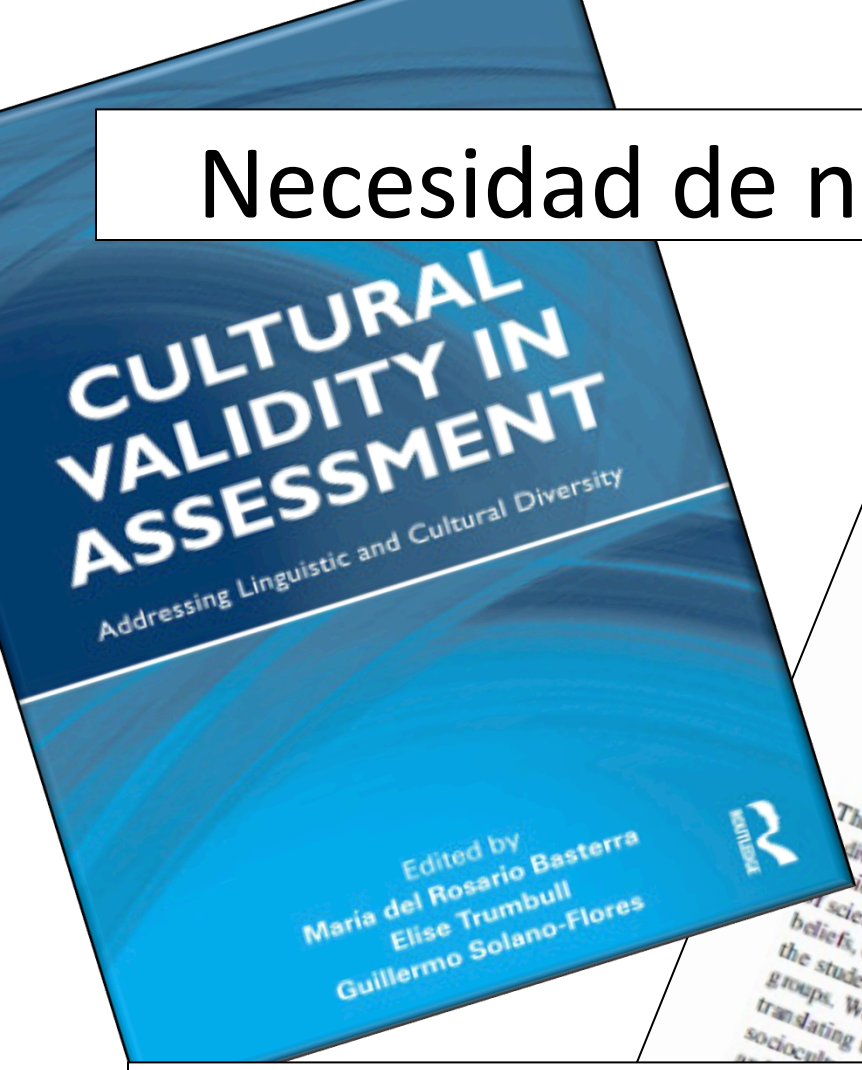


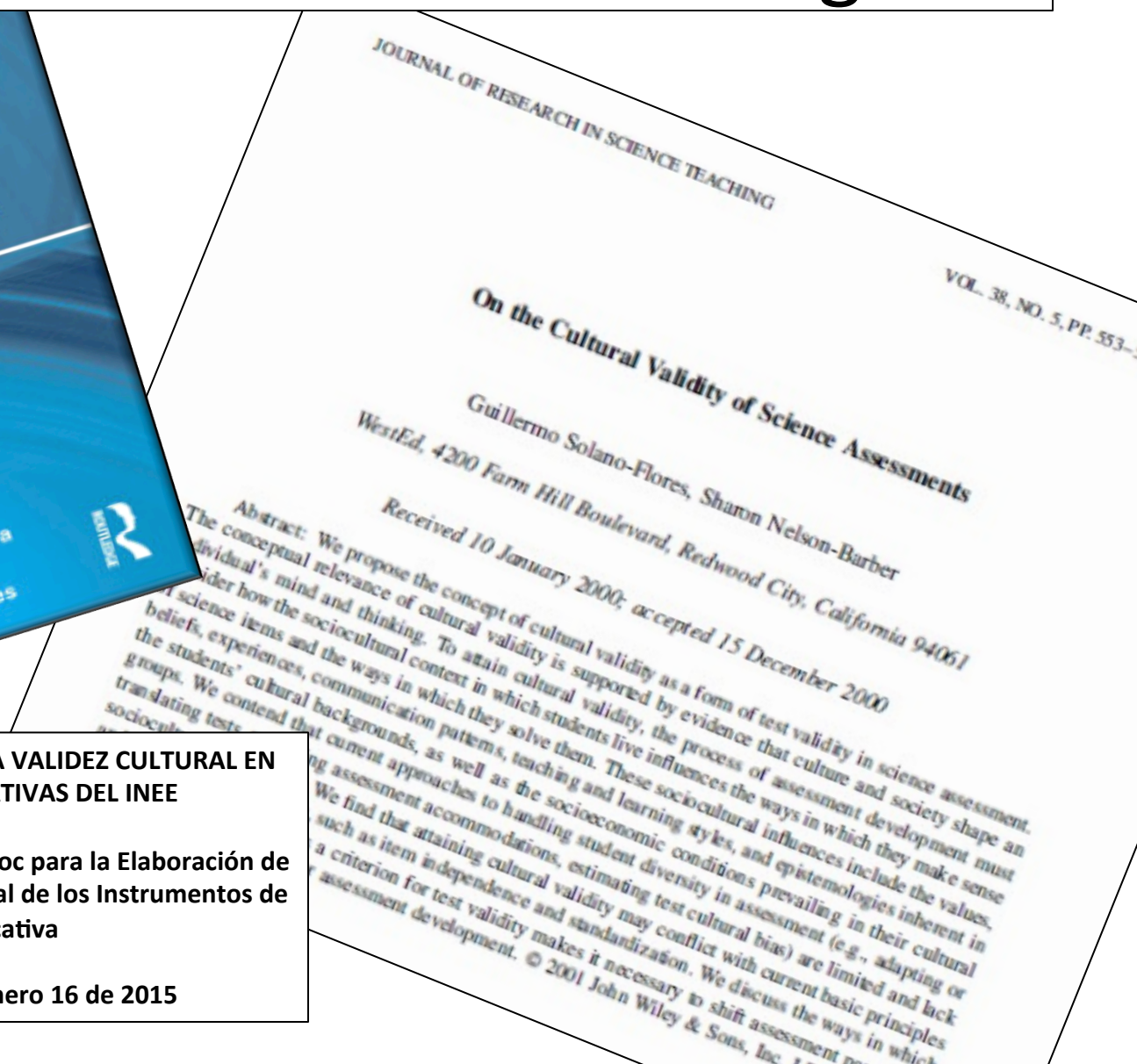
Adaptado de: Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Limitaciones en el análisis del funcionamiento diferencial de ítems

- Supone heterogeneidad en las poblaciones
- Requiere de un número relativamente grande de estudiantes en cada grupo
- Se tiene que realizar separadamente por cada ítem
- Se identifica a los ítems inadecuados muy tarde en el proceso de evaluación
- Conduce a pensar en validez en términos de un grupo poblacional, no de los constructos evaluados

Necesidad de nuevas metodologías




JOURNAL OF RESEARCH IN SCIENCE TEACHING

VOL. 38, NO. 5, PP. 553–

On the Cultural Validity of Science Assessments

Guillermo Solano-Flores, Sharon Nelson-Barber
WestEd, 4200 Farm Hill Boulevard, Redwood City, California 94061

Received 10 January 2000; accepted 15 December 2000

Abstract: We propose the concept of cultural validity as a form of test validity in science assessment. The conceptual relevance of cultural validity is supported by evidence that culture and society shape an individual's mind and thinking. To attain cultural validity, the process of assessment development must consider how the sociocultural context in which students live influences the ways in which they make sense of science items and the ways in which they solve them. These sociocultural influences include the values, beliefs, experiences, communication patterns, teaching and learning styles, and epistemologies inherent in the students' cultural backgrounds, as well as the socioeconomic conditions prevailing in their cultural groups. We contend that current approaches to handling student diversity in assessment (e.g., adapting or translating tests ... assessment accommodations, estimating test cultural bias) are limited and lack ... We find that attaining cultural validity may conflict with current basic principles ... such as item independence and standardization. We discuss the ways in which ... a criterion for test validity makes it necessary to shift assessment ... assessment development. © 2001 John Wiley & Sons, Inc.

PROMOCIÓN Y EVALUACIÓN DE LA VALIDEZ CULTURAL EN LAS ACTIVIDADES EVALUATIVAS DEL INEE

Consejo Técnico Especializado Ad-Hoc para la Elaboración de Criterios Técnicos de Validez Cultural de los Instrumentos de Evaluación Educativa

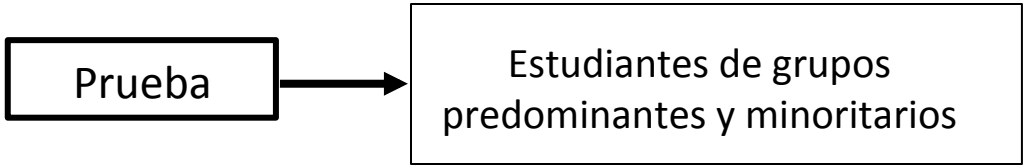
Ciudad de México, D.F., Enero 16 de 2015

Definiciones de validez

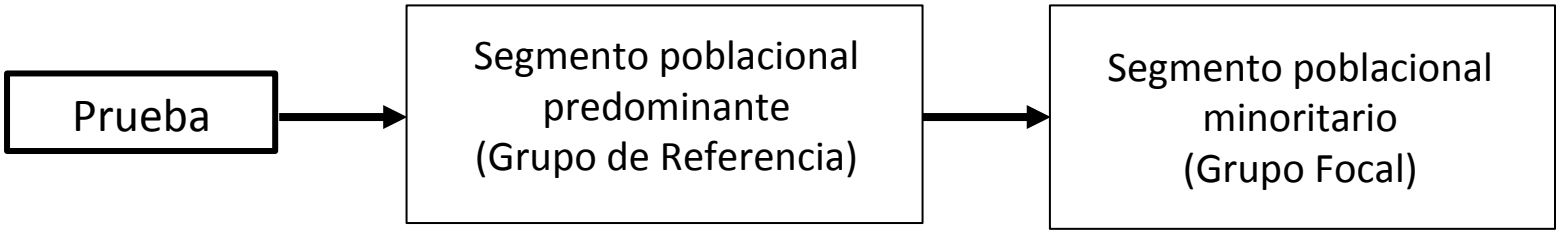
- …juicio evaluativo integral sobre el grado en que la evidencia empírica y el razonamiento teórico apoyan la adecuación y la propiedad de las inferencias y las acciones basads en calificaciones de pruebas u otras formas de evaluación (Messick, 1989)
- …grado en que la evidencia y teórica apoyan a las interpretaciones de las calificaciones en pruebas para los usos propuestos de esas pruebas (AERA, APA, & NCME, 2014)
- …grado en el que se pueden hacer generalizaciones apropiadas acerca de las habilidades y el conocimiento de los estudiantes en un área de dominio determinada con base en su desempeño en una prueba

Argumentos y evidencia en apoyo a generalizaciones e interpretaciones

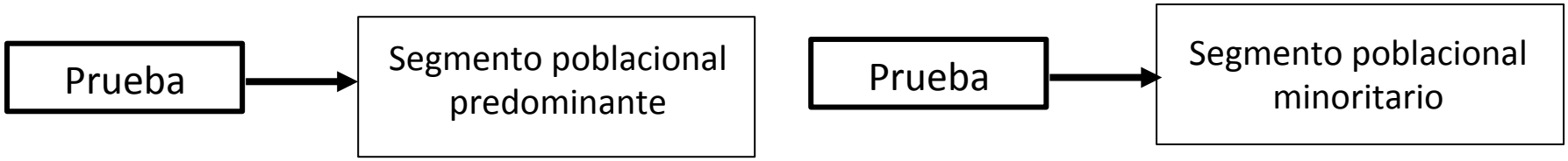
Población total



Grupo de referencia y grupo focal



Desagregación



Ejemplo 1: Desagregegación

Número mínimo de ítems necesarios para obtener coeficientes de generalizabilidad aceptables

Grupo Predominante	Grupo Minoritario 1	Grupo Minoritario 2
16	19	25

Basado en: Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice 25*(1), 13-22.

Ejemplo 2: Entrevistas cognitivas y validación cultural

Generalizability of Cognitive Interview-Based Measures Across Cultural Groups

Guillermo Solano-Flores, *University of Colorado,*
and Min Li, *University of Washington*

We addressed the challenge of scoring cognitive interviews in research involving multiple cultural groups. We interviewed 123 fourth- and fifth-grade students from three cultural groups to probe how they related a mathematics item to their personal lives. Item meaningfulness—the tendency of students to relate the content and/or context of an item to activities in which they are actors—was scored from interview transcriptions with a procedure similar to the scoring of constructed-response tasks. Generalizability theory analyses revealed a small amount of score variation due to the main and interaction effect of rater but a sizeable magnitude of measurement error due to the interaction of person and question (context). Students from different groups tended to draw on different sets of contexts of their personal lives to make sense of the item. In spite of individual and potential cultural communication style differences, cognitive interviews can be reliably scored by well-trained raters with the same kind of rigor used in the scoring of constructed-response tasks. However, to make valid generalizations of cognitive interview-based measures, a considerable number of interview questions may be needed. Information obtained with cognitive interviews for a given cultural group may not be generalizable to other groups.

Keywords: assessment, cognitive interviews, cognitive validity, culture, generalizability theory

Research in testing is increasingly using students, inferred cognitive activity to validate tests. Cognitive validation is based on examining student verbalizations and responses to interview questions with the purpose of determining the extent to which an instrument elicits the knowledge and reasoning it is intended to tap (Ayala, Shavelson, Yin, & Schultz, 2002; Baxter, Elder, & Glaser, 1996; Baxter & Glaser, 1998; Hamilton, Nussbaum, & Snow, 1997; Huberman & Levine, 1999; Karabenick et al., 2006; Megone, Cai, Silver, & Wang, 1994; Nuthall & Alton-Lee, 1995; Ruiz-Primo, Shavelson, Li, & Schultz, 2001).

In spite of the valuable contribution of cognitive-based interview approaches to test validation, two issues need to be properly addressed before they can become a part of standard practices in testing. First is the challenge of reliably coding or scoring interviews. Rater inconsistency may result from the intrinsic ambiguity of speech (van Geert & van Dijk, 2003), variability in the length and discursive structure of verbalizations (Heath, 1983), and the interaction between the complexity of the interview questions and the complexity of the cognitive processes assessed (Welzel & Roth, 1998). Unfortunately, with few exceptions (e.g., Ferrara et al., 2004; Ruiz-Primo, Shavelson, Li, & Schultz, 2001), the procedures used to code student verbalizations and interviews in test validation are rarely reported in detail.

The second issue derives from the fact that most of the theoretical developments in cognitive theory are based on findings from studies conducted mainly with middle-class, white students. The extent to which those findings can be generalized to other groups remains uncertain (see Pellegrino, Chudowsky, & Glaser, 2001, p. 300). Cultural groups may vary considerably on what is customary and socially acceptable in aspects of verbal interaction, which are relevant in the classroom (see Delpit, 1995; Estrin & Nelson-Barber, 1995; Lee, 2002). One of these aspects is the length of time persons wait before they respond to a question (see Chi, Feltovich, & Glaser, 1981) and which may be reflected in the time students from different cultures take to think about questions they are asked in interviews. For example, in some cultures, children are discouraged to respond rapidly to questions—a behavior that is considered impolite (Kusimo et al., 2000). Also, cultural groups may have different communication styles that may influence, among other things, the length of utterances in a conversation, the pause length between utterances, or the level of detail of information expected when individuals ask or respond to questions (Heath, 1983; Stockwell, 2002). Moreover, due to different epistemologies, individuals from different cultural groups may have different ways of interpreting questions

Guillermo Solano-Flores, School of Education, University of Colorado, Boulder-Education, 249 UCB, Boulder, CO 80309; Guillermo.Solano@colorado.edu.

Table 4. EVC and Rounded Percentage of Score Variation for the Three Cultural Groups Separately and Combined: Random p × q × r Model Assuming one Question and one Rater

	Group S ($n = 27$)		Group E ($n = 55$)		Group T ($n = 41$)		Combined ($n = 123$)	
Source	EVC	Percent	EVC	Percent	EVC	Percent	EVC	Percent
p (person)	0.012	6	0.015	5	0.015	6	0.025	9
q (question)	0.025	12	0.019	6	0.023	9	0.022	8
r (rater)	0.001	0	0.001	0	0.005	2	0.002	1
pq	0.105	51	0.168	55	0.129	49	0.139	50
pr	0.000	0	0.011	4	0.005	2	0.007	3
qr	0.004	2	0.005	2	0.008	3	0.003	1
pqr,e	0.061	29	0.085	28	0.076	29	0.078	28
ρ^2	0.069		0.053		0.068		0.100	

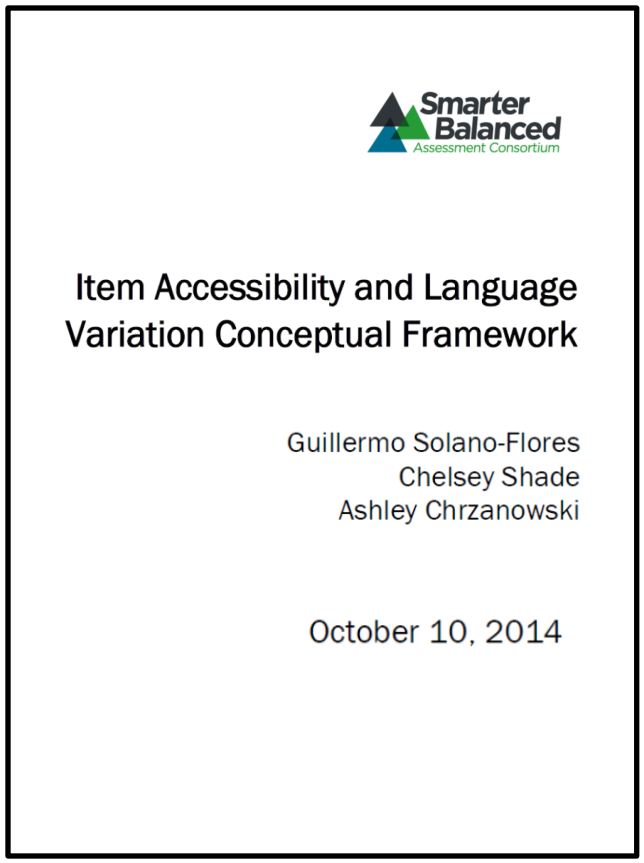
Table 5. Descriptive Statistics and Comparison of Item Meaningfulness Scores Across Groups

Context	Group S ($n = 27$)	Group E ($n = 55$)	Group T ($n = 41$)	Significance[a]	Effect Size[b]
Out of school	2.578 (0.247)	2.311 (0.442)	2.325 (0.358)	.011 ($F_{2,100} = 4.77$)	.086
Having fun at school	2.229 (0.566)	1.934 (0.532)	2.031 (0.419)	.069 ($F_{2,100} = 2.75$)	.044
In the classroom	2.188 (0.263)	2.096 (0.293)	1.986 (0.203)	.012 ($F_{2,100} = 4.60$)	.059
Any traditions	2.374 (0.330)	1.901 (0.676)	1.957 (0.615)	.004 ($F_{2,100} = 5.76$)	.100

[a]Due to missing values, the degrees of freedom do not correspond to the sample sizes. Similar effect sizes with slightly different F values were obtained with simple one-way ANOVAs.
[b]Effect size was estimated using partial ρ^2.

Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice, 28* (2), 9-18.

Ejemplo 3: Sistematización en el desarrollo de pruebas



Item Accessibility and Language Variation Conceptual Framework

Guillermo Solano-Flores
Chelsey Shade
Ashley Chrzanowski

October 10, 2014

https://www.smarterbalanced.org/wp-content/uploads/2015/08/Item-Accessibility-and-Language-Variation-Conceptual-Framework.pdf

Contexto social implícito en el diseño de ítems

El tópico o la situación refleja la vida y el ambiente social de…		
…muchos alumnos, independientemente de su grupo cultural, étnico o socioeconómico.	…individuos de un grupo cultural, étnico o socioeconómico específico pero es familiar o conocido por individuos de muchos otros grupos.	…individuos de un grupo cultural, étnico o socioeconómico específico y es poco familiar o es desconocido por individuos de otros grupos.

Adaptado de: **Solano-Flores, G.**, Shade, C., & Chrzanowski, A. (2014). *Item accessibility and language variation conceptual framework. Submitted to the Smarter Balanced Assessment Consortium.* October 10. http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/11/ItemAccessibilityandLanguageVariationConceptualFramework_11-10.pdf

Ejemplo 4: Sistematización en el diseño de ítems

Nombres de personajes ficticios a usar en las pruebas desarrolladas por un sistema de evaluación

Personajes Femeninos		Personajes Masculinos	
Destiny	*Rachel*	*Tyler*	*Eli*
Kiara	*Rebecca*	*Brandon*	*Abraham*
Alissa	*Sarah*	*Christian*	*David*
Yuki	*Diana*	*Hiro*	*Carlos*
Indira	*Maria*	*Manzur*	*Eduardo*
Meyumi	*Rosa*	*Cheng*	*Roberto*
Molly	*Aida*	*Steve*	*Abdul*
Claire	*Adela*	*John*	*Yousef*
Emily	*Shakira*	*Mike*	*Ahmed*

Adaptado de: **Solano-Flores, G.**, Shade, C., & Chrzanowski, A. (2014). *Item accessibility and language variation conceptual framework. Submitted to the Smarter Balanced Assessment Consortium.* October 10. http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/11/ItemAccessibilityandLanguageVariationConceptualFramework_11-10.pdf

Criterios de validez cultural

- ***Marco conceptual de los instrumentos de evaluación***
- ***Especificación de poblaciones y unidades de análisis***
- ***Estrategia para considerar diversidad cultural, lingüística y socioeconómica***
- ***Especificación de ítems***
- ***Profesionales involucrados en el desarrollo de los ítems***
- ***Representación de poblaciones diversas en las muestras***
- ***Validación cognitivo-cultural***
- ***Revisión de expertos***
- ***Análisis de sesgo cultural***
- ***Estudios de generalizabilidad***
- ***Tiempos y calendarios***
- ***Mecanismos de corrección***

Enfasis en el proceso de desarrollo de pruebas: Acciones inmediatas

- Desarrollar un documento con el marco poblacional que defina grupos de interés
  - Localidad: rural y urbana
  - Niveles socioeconómicos: alto, medio y bajo
  - Etnicidad: indígena, no indígena
  - Lengua materna: español, lengua indígena
- Probar borradores de los ítems con muestras de estudiantes pertenecientes a esos grupos de interés
- Inclusión de lingüistas y antropólogos en los equipos de autores de ítems
- Inclusión de maestros provenientes de múltiples contextos sociales

Gracias!

gsolanof@Stanford.edu