PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2015, 52(2), 55-78

Pontificia Universidad Católica de Chile
Facultad de Educación

Construyendo una zona de intercambio productiva en la investigación y práctica de la evaluación educacional

Building a Productive Trading Zone in Educational Assessment Research and Practice

William P. Fisher Jr. y Mark Wilson

Graduate School of Education, University of California, Berkeley, EE.UU.

Resumen

Existen puntos de vista marcadamente distintos en muchos contextos, desde la ciencia, pasando por las aulas, hasta el mercado. Lo que podría ser sorprendente es que pueden hallarse tanto disonancias divergentes como armonías convergentes en sistemas productivos del mundo real. El valor de los resultados de evaluación generalizables depende de que estos sean innovadores a la vez que estandarizados. Estas tensiones aparentemente opuestas pueden reconciliarse en términos de objetos de demarcación, entidades compartidas por distintas comunidades que los usan y los ven de modo bastante diferente. Más aún, desde hace largo tiempo se ha considerado que la ciencia ocurre dentro de un continuo que va desde el pensar y el actuar cotidiano a la lógica y a los métodos formales; por lo tanto, no debiera ser una sorpresa que este rango se manifieste también en la investigación psicométrica. Describimos métodos y resultados en los cuales ejemplares modelados psicométricamente, conocidos como mapas de constructo y mapas Wright, funcionan como objetos de demarcación y sirven de base para analogías productivas en la evaluación educacional de las siguientes formas: (a) preservando estructuras relacionales, (b) generando mapeos isomorfos entre sistemas y (c) facilitando la sistematicidad, comprendida como sistemas de mapeo de estructuras relacionales de orden superior (Nersessian & Chandrasekaran, 2009). En este contexto conceptual, presentamos una aplicación del Sistema de Evaluación BEAR [BEAR Assessment System] y su software asociado, los que facilitan la traducción de estructuras relacionales entre sistemas, apoyando así las alianzas prácticas entre la enseñanza, la legislación, la evaluación y desarrollo curriculares, la psicometría y las tecnologías de la información (TI).

Palabras clave: mapeo de constructos, Teoría del Actor-Red, redes de traducción, modelos Rasch, objetos de demarcación

Correspondencia a:
William P. Fisher Jr.
BEAR Center, Graduate School of Education
University of California Berkeley
2000 Center Street, Suite 301, Berkeley CA 94704, USA.
Correo electrónico: wfisher@berkeley.edu

ISSN: 0719-0409 DDI: 203.262, Santiago, Chile
doi: 10.7764/PEL.52.2.2015.4

Abstract

Markedly diverse viewpoints are in play in many varied contexts, from science to the classroom to the marketplace. Perhaps surprisingly, both divergent dissonances and convergent harmonies are routinely found together in productive real-world systems. The value of generalizable assessment outcomes hinges on their being both innovative and standardized. These apparently opposite tensions can be reconciled in terms of boundary objects, entities shared by different communities that use and view them quite differently. Further, science has long been seen as taking place on a continuum from everyday thinking and acting to formal logic and methods, so it should not be surprising to find this range manifest as well in psychometric research. We describe methods and results in which psychometrically modeled exemplars known as construct maps and Wright maps function as boundary objects and serve as a basis for productive analogies in educational assessment by (a) preserving relational structures, (b) making isomorphic mappings between systems, and (c) facilitating systematicity, understood as mapping systems of higher order relational structures (Nersessian & Chandrasekaran, 2009). In this conceptual context, we present an application of the BEAR Assessment System and its accompanying software, facilitating translations of relational structures across systems in support of practical alliances of teaching, policy-making, assessment and curriculum development, psychometrics, and information technology (IT).

Keywords: construct mapping, Actor Network Theory, translation networks, Rasch models, boundary objects

La psicología y las ciencias sociales están marcadas por su falta de consenso en cuanto a métodos y estándares de medición y, no coincidentemente, por un enfoque superficial tipo «recetas de cocina» de aplicaciones de medición. Los trabajos publicados durante varias décadas (Bakker, van Dijk, & Wicherts, 2012; Berkson, 1938; Bolles, 1962; Coats, 1970; Cohen, 1994; Guttman, 1985; Michell, 1999; Roberts, 1994; Taagepera, 2008; Wilson, 1971) documentan rangos de variación alarmantemente amplios en cuanto a los métodos considerados aceptables, desacuerdos con respecto a conceptos básicos, y una amplia indiferencia frente a las ventajas científicas y prácticas de los enfoques que requieren pruebas experimentales y predicciones teóricas en lugar de suposiciones no examinadas y descripciones puramente empíricas. La importancia de la medición se reconoce universalmente como un fundamento, y métodos con bases sólidas y practicables han estado disponibles durante décadas en la literatura de investigación, en los textos de estudio y en software. Sin embargo, en la práctica, es poco común encontrar pruebas, evaluaciones o encuestas diseñadas con teorías explicativas del constructo medido, calibrado en unidades invariantes con incertidumbres estimables e informadas por interpretaciones cualitativas de patrones de respuesta, tanto consistentes como inconsistentes. La impresión general, bastante desalentadora, que se obtiene al revisar la literatura es que, aunque se ha llevado a cabo una gran cantidad de buen trabajo, incluyendo pruebas matemáticas, enseñanza inspiradora, software de fácil acceso, métodos simples y prácticos, resultados empíricos reproducidos y explicaciones teóricas persuasivas, la investigación en su conjunto no ha sido efectiva en eliminar los ya mencionados desacuerdos fundamentales relacionados con la medición en las ciencias sociales, ni tampoco ha logrado promover un marco que encarne las tres características importantes que se necesitan (también mencionadas anteriormente).

Tal vez una forma de acercarse a una respuesta sea mirar más allá de los límites tradicionales de la «medición» y considerar el contexto más amplio dentro del cual operan las mediciones. En este sentido, es relevante tener en cuenta que, durante varias décadas, los estudios históricos y sociales de la ciencia (Galison, 1999; Hutchins, 1995, 2014; Latour, 1987, 1993a; Nersessian, 2006, 2012) han revivido la pregunta que Hayek (1948, p. 54) formulara como «la pregunta central de todas las ciencias sociales: ¿Cómo puede la combinación de fragmentos de conocimiento existente en mentes distintas entregar resultados que, para generarse deliberadamente, requerirían que la mente directiva poseyera un conocimiento no disponible para ningún individuo por sí solo?» Hayek (1948), en conjunto con la observación de Whitehead (1911, p. 61) tomaron este cuestionamiento, señalando que la civilización no avanza mediante el pensamiento original tanto como lo hace por medio de tecnologías que permiten que las personas ejecuten exitosamente operaciones que no comprenden y que no podrían cumplir por sí solas. Por ejemplo, ninguna persona por sí sola es capaz de fabricar un automóvil, traer luz eléctrica a

un hogar o hacer que funcione un mercado de *commodities*. Cada uno de estos avances surge sin un único director, sino que mediante los esfuerzos combinados de personas expertas en campos muy variados, desde trabajadores, proveedores, oficinistas y consumidores, hasta científicos, ingenieros y matemáticos, hasta financistas y economistas, hasta educadores y legisladores. Las descripciones detalladas sobre cómo estos grupos tan diversos se han coordinado y alineado sus actividades en aplicaciones productivas de la ciencia y la tecnología (Galison, 1997; Latour, 1993a; Miller & O'Leary, 2007; Star & Griesemer, 1989) sugieren nuevas posibilidades para mejorar la calidad de los métodos empleados en psicología y ciencias sociales (Fisher, 2000, 2005, 2009; Fisher & Stenner, 2011, 2013a).

Objetos de demarcación para la medición: el mapa de constructo y el mapa Wright

Una forma en la cual la historia contemporánea y la filosofía de la ciencia encuadran estos problemas es en términos de *zonas de intercambio* (Galison, 1997, 1999; Galison & Stump, 1996) y de los *objetos de demarcación* implicados en las redes de traducción (Star & Griesemer, 1989; Woolley & Fuchs, 2011) dentro de tales zonas. Las zonas de intercambio son foros en los cuales puede realizarse un intercambio seguro de ideas entre personas de manera similar a como se hace en las áreas comerciales neutrales que, según observan los etnógrafos, emergen entre vecinos no amistosos que poseen productos que desean comprar y vender. Los dos grupos suelen dotar a los objetos intercambiados de significados y valores completamente diferentes. Dichos objetos, puesto que residen en los límites entre grupos distintos, se conocen como objetos de demarcación y se definen en términos de traducciones basadas en analogías estructuradas sistemáticamente.

Esta perspectiva amplia en la investigación en ciencias sociales sugiere que deberíamos formular la pregunta de cómo se vería un enfoque de este tipo en la investigación psicométrica; en otras palabras, ¿qué objetos de demarcación serían adecuados para sustentar una amplia gama de participantes, como los mencionados anteriormente, en la evaluación educacional? Dichos participantes podrían ser profesores, desarrolladores de evaluaciones, psicometristas, expertos en tecnologías de la información (TI), desarrolladores curriculares, responsables de formular políticas y otros. En este artículo, describimos un enfoque específico para establecer una zona de intercambio, empleando objetos de demarcación, los que se basan en los conceptos fronterizos del *mapa de constructo* (Wilson & Sloane, 2000) y el *mapa Wright* (Wright & Stone, 1979), junto con una forma de organizarlos conocida como *modelamiento de constructos* (Wilson, 2005).

En primer lugar, un mapa de constructo entrega una representación concreta de las expectativas teóricas (idealmente sostenidos por resultados empíricos) sobre la naturaleza del constructo en una evaluación o encuesta. Los mapas de constructo determinan puntos particularmente útiles dentro de un continuo del constructo y entregan una definición coherente y sustantiva del contenido de la evaluación (Wilson, 2005). Un mapa de constructo es un ordenamiento bien pensado, teóricamente justificado y adecuadamente investigado de puntos de desempeño cualitativamente diferentes enfocados en una característica. Los mapas de constructo derivan en parte de la investigación enfocada en la estructura subyacente del dominio y en parte de juicios profesionales sobre qué constituye los niveles superiores o inferiores de desempeño o competencia, pero también están informados por estudios empíricos sobre cómo piensan las personas y cómo actúan en la práctica (National Research Council, Division of Behavioral and Social Sciences and Education, Center for Education, & Mathematics Learning Study Committee, 2001).

La Tabla 1 muestra un ejemplo de un mapa de constructo, desarrollado como parte de un proyecto de evaluación de Química en la Universidad de California, Berkeley, llamado «Perspectives of Chemists» [Perspectivas de los químicos] (Claesgens, Scalise, Wilson, & Stacy, 2009). El proyecto intentó encarnar la comprensión de la química desde un nivel de sofisticación que va de principiante a avanzado, por medio de un conjunto de mapas de constructo. La Tabla 1 presenta el aspecto *Materia,* que se ocupa de describir las visiones atómica y molecular de la misma. Debe tenerse en cuenta, sin embargo, que esta no es una descripción de contenido típica, como la que se halla en la mayoría de los textos de estudio, sino que describe cómo progresa la visión sobre la materia de un estudiante desde una perspectiva continua, propia del mundo real, a una particulada, y cómo luego va aumentando en sofisticación. Existen muchas formas de exponer un mapa de constructo, y esta incluye casi todos los componentes típicos. De izquierda a derecha, estos son: un nombre para cada nivel, luego un resumen

del contenido en el que se enfocan los estudiantes en dicho nivel, seguido por una descripción del pensamiento de los estudiantes en este nivel y, finalmente, por ejemplos de ítems que podría pedírsele completar a un estudiante en este nivel. Algunos mapas van más allá de esto e incluyen además ejemplos de respuestas de los estudiantes a los ítems en cada nivel. Claramente, la existencia de un mapa de constructo satisface la primera parte de nuestros tres criterios: que el diseño incluya teorías explicativas del constructo. Sin embargo, esto no es suficiente por sí solo, y ni siquiera es toda la primera parte, ya que un mapa de constructo por sí mismo no entrega una medida del constructo. Esto será representado mediante el otro objeto de demarcación, el mapa Wright.

Tabla 1
Marco de perspectivas de los químicos; variable *materia*

Nivel de éxito	Ideas grandes	Descripciones del nivel	Ejemplares de ítems
10-12 Construcción ¿Por qué la composición, estructura, propiedades y cantidades? (Uso de modelos)	La composición, estructura y propiedades de la materia se explican por medio de las distintas fuerzas de las interacciones entre partículas (electrones, núcleos, átomos, iones, moléculas) y los movimientos de estas partículas.	Los estudiantes son capaces de razonar empleando modelos químicos normativos, utilizan estos modelos para explicar y analizar la fase, composición y propiedades de la materia. Emplean modelos químicos precisos y apropiados en sus explicaciones, y comprenden los supuestos usados para construir los modelos.	a) Composición: ¿Cómo podemos explicar la composición? b) Estructura: ¿Cómo podemos explicar la estructura tridimensional? (por ejemplo, estructura cristalina, formación de gotas) c) Propiedades: ¿Cómo podemos explicar las variaciones en las propiedades de la materia? (por ejemplo, punto de ebullición viscosidad, dureza, pH, etc.) d) Cantidad: ¿Qué supuestos hacemos cuando medimos la cantidad de materia? (por ejemplo, ley de los gases no ideales, masa promedio)
7-9 Formulación ¿Cómo podemos pensar sobre las interacciones entre moléculas? (Multi-relacional)	La composición, estructura y propiedades de la materia están relacionadas con la manera en que los electrones se distribuyen entre los átomos.	Los estudiantes están desarrollando una comprensión más coherente que indica que la materia está hecha de partículas y que el ordenamiento de dichas partículas se relaciona con las propiedades de la materia. Sus definiciones son precisas, pero su comprensión no está totalmente desarrollada, por lo tanto, el razonamiento de los estudiantes se limita a mecanismos causales en vez de explicativos. En sus interpretaciones de situaciones nuevas, es posible que los estudiantes sobregeneralicen al intentar relacionar múltiples ideas y fórmulas de constructos.	a) Composición: ¿Por qué la tabla periódica es una hoja de ruta para los químicos? (¿Por qué es una tabla «periódica»?) ¿Cómo podemos pensar los ordenamientos de los electrones en los átomos? (por ejemplo, capas, orbitales) ¿Cómo se relacionan los números de electrones de valencia con la composición? (por ejemplo, transferencia o intercambio). b) Estructura: ¿Cómo explica la estructura tridimensional las conexiones entre átomos (enlaces) y los movimientos de los átomos? (el diamante es rígido, el agua fluye, el aire es invisible). c) Propiedades: ¿Cómo puede clasificarse la materia de acuerdo a los enlaces? (los sólidos iónicos se disuelven en agua, sólidos covalentes duros, fase de la materia). d) Cantidad: ¿Cómo puede relacionarse una cantidad de materia con otra? (por ejemplo, masa/moles/número, ley de los gases ideales, ley de Beer).
4-6 Reconocimiento ¿Cómo describen	La materia se categoriza y se describe de	Los estudiantes exploran el lenguaje y los símbolos empleados por los químicos	a) Composición: ¿Cómo muestra tendencias la tabla periódica? ¿Cómo se clasifican mediante letras y

la materia los químicos? (Unirrelacional)	acuerdo a varios tipos de partículas subatómicas, átomos y moléculas.	para describir la materia. Relacionan el número de electrones, protones y neutrones con los elementos y la masa, y el ordenamiento y los movimientos de los átomos con la composición y la fase. Los modos de pensar sobre la materia se limitan a relacionar una idea con otra en un nivel de comprensión simple.	símbolos los elementos, compuestos y mezclas? b) Estructura: ¿Cómo difieren los ordenamientos y movimientos de los átomos en los sólidos, líquidos y gases? c) Propiedades: ¿Cómo pueden predecirse propiedades utilizando la tabla periódica? d) Cantidad: ¿Cómo calculan los químicos las cantidades de partículas? (por ejemplo, número, masa, volumen, presión, mol)
1-3 Nociones ¿Qué sabes sobre la materia?	La materia tiene masa y ocupa espacio. Puede clasificarse de acuerdo a cómo ocupa espacio.	Los estudiantes articulan ideas sobre la materia y usan la experiencia, la observación y el razonamiento lógico para entregar evidencia. Su enfoque es principalmente macroscópico (no particulado).	a) Composición: ¿Cómo difiere la materia de la energía, los pensamientos y los sentimientos? b) Estructura: ¿Cómo difieren entre sí los sólidos, los líquidos y los gases? c) Propiedades: ¿Cómo puedes usar las propiedades para clasificar la materia? d) Cantidad: ¿Cómo puedes medir la cantidad de materia?

En segundo lugar, para que un mapa de constructo, que es una expresión de una intención, sea una forma útil de medición, debe fortalecerse empíricamente, produciendo otra versión del mapa conocida como mapa Wright. El apoyo empírico requiere del desarrollo de (a) un conjunto de ítems que encarnen el constructo en términos de las respuestas de una persona a los ítems; (b) un plan para transformar esas respuestas en datos; y (c) un método de calibración de un instrumento para esas respuestas a los ítems, para así poder usarlas como la representación empírica del constructo. En el enfoque específico de la iniciativa que describimos, la representación empírica del constructo se denomina mapa Wright.

La Figura 1 presenta un mapa Wright correspondiente al mapa de constructo mostrado en la Tabla 1. El lado izquierdo de este mapa muestra la distribución medida de estudiantes que respondieron a los ítems sobre *Materia,* mientras que el lado derecho indica la dificultad calibrada de un subconjunto de seis de las tareas. La estimación de estas ubicaciones de estudiantes e ítems se basó en el modelo Rasch (Rasch, 1960), el cual entrega el mapeo tanto de los ítems como de los estudiantes dentro la misma escala logit. Como se puede observar, las respuestas de los estudiantes se separaron en dos segmentos distintos de la escala logit, lo que facilitó ubicar a los estudiantes por debajo (aproximadamente) de -0.05 logits en el nivel 1 del mapa de constructo, es decir, para el nivel de *Nociones,* y a los estudiantes por sobre esa línea (destacados) en el nivel 2 de este mapa de constructo, el nivel de *Reconocimiento.* Para este grupo particular de estudiantes, solo los dos primeros niveles del mapa se mantuvieron. El análisis también genera estadísticas de ajuste que permiten marcar vectores de respuesta inconsistentes y otros índices que muestran cuán bien se ajustan a los datos los niveles especificados por el modelo. Se generan tablas de coeficientes de confiabilidad y de errores estándar, en donde pueden también realizarse comparaciones interobservador. El mapa Wright es el resultado del exitoso escalamiento Rasch del constructo; esto, combinado con el mapa de constructo, satisface las otras partes de nuestro trío de características importantes: que el instrumento se diseñe con teorías explicativas del constructo medido, que se calibre en unidades invariantes con incertidumbres estimables y que esté informado por interpretaciones cualitativas de patrones de respuesta tanto consistentes como inconsistentes. Es importante tener en cuenta que el exitoso escalamiento Rasch del constructo es por sí mismo insuficiente: se satisfacen las tres características solo cuando los dos objetos de demarcación están en sincronía (como se ilustró en la Tabla 1).

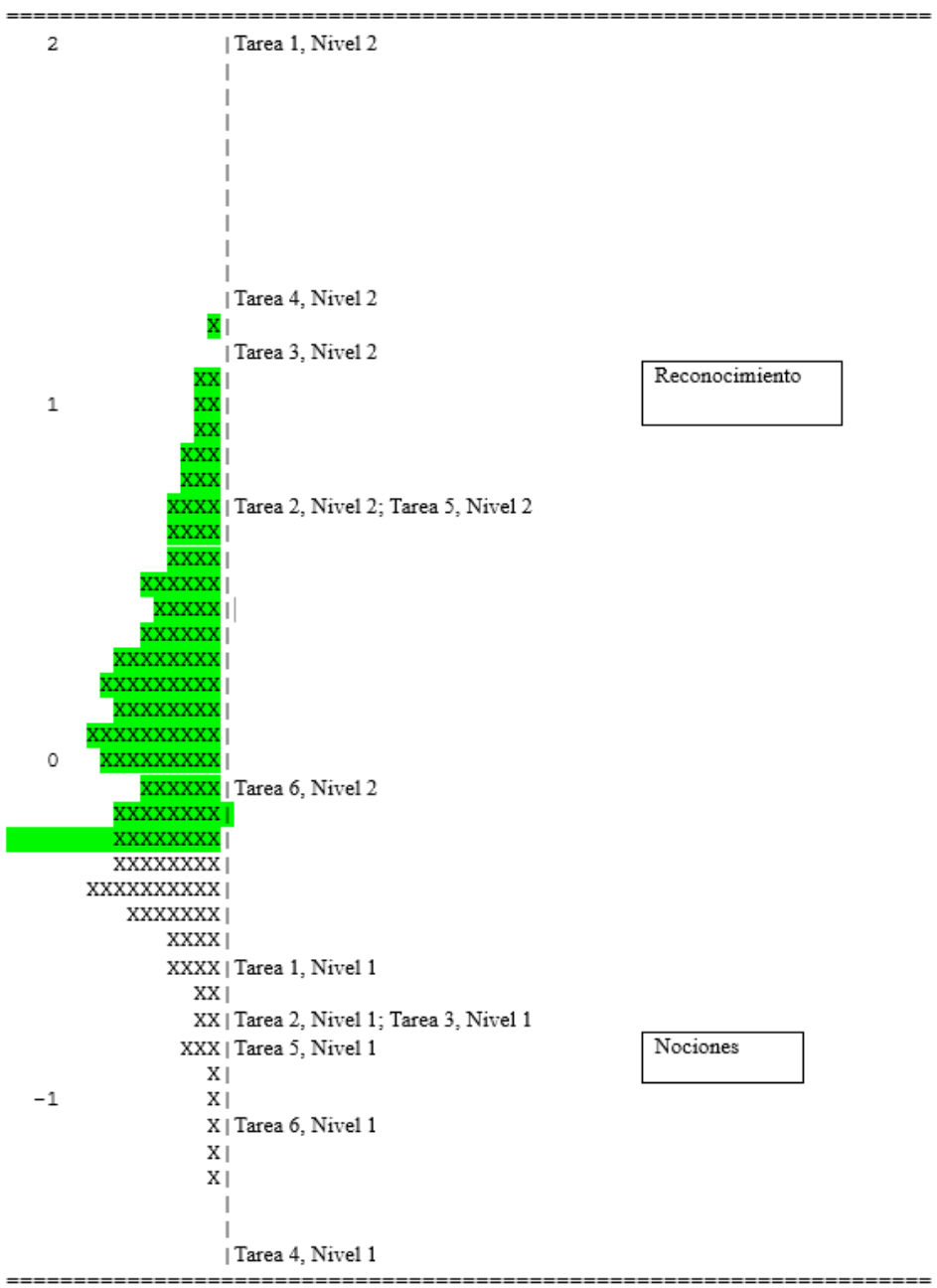

Figura 1. Mapa Wright de la variable Materia en el instrumento ChemQuery (Crédito Parcial, Umbrales generalizados de ítems). *Los números de la izquierda están en una unidad de medida denominada «logit», o logaritmo natural de la razón de probabilidades, en la cual valores mayores indican mejor desempeño de los estudiantes.

En tercer lugar, estas dos representaciones están coordinadas conjuntamente dentro de un ciclo conocido como modelamiento de constructo, ilustrado en la Figura 2. En esta figura existen dos pasos intermedios entre el mapa de constructo y el mapa Wright, específicamente el *diseño de los ítems* y el *espacio de resultados,* que corresponden a: (a) el diseño de los ítems que se espera produzcan respuestas posibles de interpretar como indicativas de niveles específicos del mapa de constructo, y (b) el esquema para valorar estas respuestas dentro de los niveles del mapa de constructo (y, posiblemente, también otras categorías). Este ciclo de desarrollo de instrumentos se itera hasta lograr una consistencia

suficiente entre las intenciones y los resultados empíricos (véase Wilson, 2005, para más detalles), luego de lo cual ya es posible investigar la evidencia de confiabilidad y validez del instrumento, y, eventualmente, usarlo de forma directa.

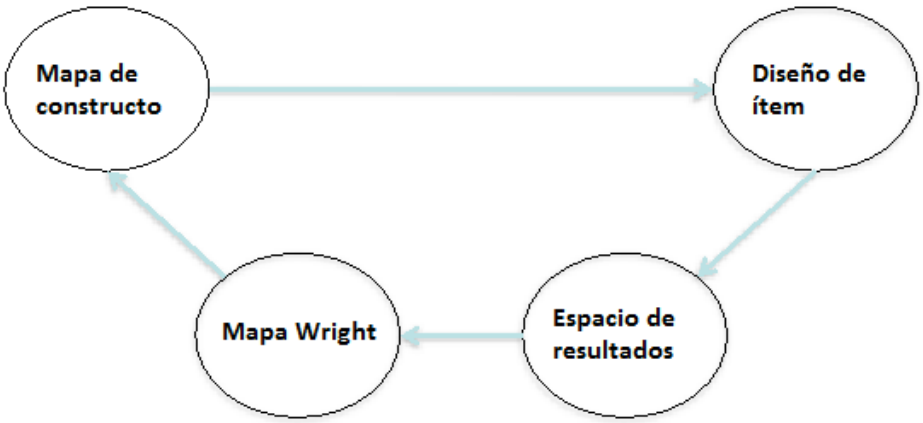

Figura 2. Modelamiento de constructos.

Imágenes como estas han demostrado ser valiosas para mejorar la confiabilidad, validez y utilidad de las mediciones psicométricas en una amplia gama de campos, desde educación (Black, Wilson, & Yao, 2011) a salud (Best, 2008; Ewert, Allen, Wilson, Üstün, & Stucki, 2010; Smith, 2005; Wilson, Allen, & Li, 2006) y psicología (Dawson, Xie, & Wilson, 2003; Kaiser & Wilson, 2000). Los éxitos a la fecha sugieren la necesidad de realizar más estudios para establecer si y cómo los mapas de constructo funcionan como objetos de demarcación en zonas de intercambio, y cómo su uso en este sentido podría expandirse y mejorarse.

Puede decirse que estos objetos de demarcación se ubican dentro de una zona de intercambio, como muestra la Figura 3, la cual ilustra ejemplos de puntos de paso y alianzas específicas para el contexto de la evaluación educacional, como se concibe en el enfoque del modelamiento de constructos. Esta es una instancia de la figura de Star y Griesemer (1989) mostrada en la Figura 4, aunque es algo más compleja, dado que permite dos objetos de demarcación en lugar de uno, en la forma de la figura de Nersessian (2012, p. 227) que ilustra las interconexiones entre los problemas de un laboratorio, los investigadores y las estrategias.

Figura 3. Una red de traducción con dos objetos frontera.

Por ejemplo, al examinar la Figura 4 podemos ver cómo los dos objetos de demarcación pueden informar el proceso completo de instrucción y evaluación. Como ejemplo de un episodio de este dominio, suponga que un profesor está usando los resultados de una evaluación para planificar el próximo paso en la instrucción. Ahora bien, el trasfondo de un evento como este es complejo, pero se

puede suponer (siguiendo de izquierda a derecha en la Figura 4) que los temas enseñados fueron decididos mediante un proceso de desarrollo de políticas llevado a cabo por los responsables de formular políticas, desarrolladores curriculares y profesores (entre otros), el cual dio como resultado un conjunto de documentos curriculares. En este proceso, los mapas de constructo (tal vez desplegados en relaciones complicadas dentro de una progresión de aprendizaje) pueden entregar una herramienta de comunicación clave para desarrolladores curriculares y profesores. Se evidencia aún más profundamente el trasfondo de la situación (observando el lado derecho de la Figura 4), considerando que las evaluaciones que el profesor podría usar para informar la instrucción (es decir, evaluaciones formativas) habían sido producidas por desarrolladores de evaluaciones, siguiendo los mapas de constructo, los cuales fueron desarrollados por expertos en evaluación y psicometristas. Con la ayuda de los psicometristas, los desarrolladores curriculares idearon un método para puntuar respuestas que es sensible tanto a potenciales concepciones erróneas como al establecimiento de pre-requisitos para instrucción individual. Por supuesto, el profesor deseará saber si los alumnos aprendieron lo que se les enseñó, y para hacerlo deberá emplear una evaluación sumativa.

Esta puede fortalecerse mediante una conexión sistemática entre las dos formas de evaluación, que es entregado por el mapa Wright. Dentro de este mapa pueden mostrarse no solo los ítems de evaluación formativa y sumativa, sino que también las posiciones de los estudiantes en estos dos conjuntos de ítems. Debe tenerse en cuenta que los responsables de formular políticas participarán en la especificación de las evaluaciones sumativas (o al menos en su motivación), dado su legítimo interés en monitorear los progresos de los estudiantes. Cada parte interesada en el proceso tiene experiencia en una o más áreas complementarias, y sabe poco o nada sobre los aspectos técnicos de las contribuciones de las demás. Incluso, si un individuo tiene múltiples áreas de experiencia, no habría suficiente tiempo ni recursos para lograr todo lo que se hace rutinariamente mediante la coordinación y el alineamiento de las actividades organizadoras del campo.

La zona de intercambio, los objetos de demarcación y las traducciones mediante analogía

En ciencia, los objetos de demarcación, muchas veces representados como imágenes de varios tipos (Daston, 2004; Daston & Galison, 1992, 2007; Dear, Hacking, Jones, Daston, & Galison, 2012; Galison, 2008; Ihde, 1998, 2012), son adaptables y traducibles a través de diferentes perspectivas. Ningún punto de vista domina a todos los demás, puesto que la efectividad de las definiciones operacionales de cada perspectiva se materializa solo en términos de la proyección social general del objeto dentro de esa perspectiva. Los científicos piensan y actúan con respecto a los objetos de demarcación en formas que no difieren cualitativamente de cómo los niños aprenden mediante el juego (Nersessian, 1996) y de cómo el pensamiento y el actuar cotidianos se relacionan con objetos conversacionales (Nersessian & Chandrasekaran, 2009). En ciencias, el juego infantil, el pensamiento cotidiano y el razonamiento se basa en modelos y se sitúa en redes distribuidas que facilitan la producción de analogías imitativas activas mediante ensayo y error (Nersessian, 2002, 2006, 2008, 2012).

¿Qué quiere decir esto para la psicología y la educación? El concepto de zona de intercambio de Galison (1999), como un área en la cual pueden traspasarse ideas y donde se puede obtener valor sin suponer una reducción a un punto de vista universal común, es una metáfora adecuada para organizar un programa positivo de investigación y práctica en el campo de la psicología y las ciencias sociales. Cada vecindario específico en una comunidad como esta puede incluir a personas que comparten una perspectiva general con respecto a un proceso, resultado o meta. Tómese como ejemplo el campo de la evaluación educacional, donde una serie de tipos significativos de participantes, como profesores, desarrolladores de evaluaciones, psicometristas, expertos en TI, desarrolladores curriculares, responsables de formular políticas y otros, tienen sus propios intereses específicos. Estos, a pesar de estar relacionados, difieren en aspectos importantes.

El problema es cómo avanzar para cumplir estos intereses de forma conjunta más efectivamente de lo que permiten los métodos actuales. Una pista significativa sobre cómo abordar este problema se halla en las dificultades experimentadas al educar a los usuarios finales, como profesores, teóricos o investigadores experimentales, sobre los aspectos técnicos específicos de la psicometría. Comunicar la complejidad de los modelos, los procesos de estimación, las evaluaciones de incertidumbre y de ajuste al modelo y la interpretación de los resultados, puede provocar un considerable nivel de frustración en

todas las partes. Por otra parte, a los psicometristas podría parecerles que los problemas experimentados por los profesores o los formulados por investigadores sustantivos son tan incomprensibles como para estos dos últimos grupos lo son los aspectos matemáticos. En este caso, lo que suele producirse en lugar de diálogos mutuamente informativos entre los distintos grupos es algo «como el juego paralelo descrito por Piaget en preescolares: estos hablan (y juegan) en compañía de los demás en vez de *hacia* y *con* los demás» (Bond & Fox, 2015, p. 299). Esta analogía con el desarrollo humano plantea la pregunta de si puede esperarse que los investigadores maduren, de modo similar a como lo hacen los preescolares cuando finalmente llegan a compartir en relaciones más recíprocas. ¿Podría haber otra forma de mediar en las relaciones entre diferentes áreas de conocimiento técnico sin esperar el tipo de maestría compartida que se exhibe cuando todos los participantes de un juego de lenguaje lo juegan con fluidez? Siguiendo la idea de Nersessian (1996) sobre cómo niños y científicos aprenden jugando mediante ensayo y error, ¿podría acaso existir una vía productiva de desarrollo para las partes interesadas en la investigación educacional que no necesariamente involucre hablar (y jugar) hacia y con los demás en mayor medida de lo que ya es normal?

La visión general de Galison (1997) sobre los cambios de paradigma filosóficos de las últimas décadas entrega pistas sobre cómo podría abordarse este problema. Históricamente, en educación y en muchos otros campos, se ha aceptado que el estatus objetivo de los datos entrega razones convincentes y una base lógica para coordinar actividades a través de las perspectivas un tanto convergentes y un tanto divergentes de los distintos grupos de partes interesadas. Como se ha demostrado abundantemente en la literatura antipositivista (Kuhn, 1970; Toulmin, 1961, 1982), este énfasis positivista o moderno en los datos, como elemento primario, rara vez se condice con la práctica real. La atención puede centrarse en los datos solo mientras se despliegue algún concepto teórico, es decir, los datos se vuelven «datos» solo a la luz de una expectativa teórica. Incluso, si no está disponible una teoría explícita, compartir y comunicar lo que se ha observado con respecto a las cosas del mundo requiere ideas y conceptos suficientemente estables como para mediar en las relaciones de forma regular y predecible (aunque tal vez revisable). La perspectiva antipositivista, por su parte, ha sido también considerada deficiente por estar atrapada en una postura relativista que enfatiza dependencias históricas, culturales y lingüísticas, lo que termina siendo contraproducente (Latour, 1991).

La perspectiva postpositivista (Galison, 1999), inmoderna [unmodern] (Dewey, 2012) o amoderna [amodern] (Latour, 1990, 1991, 1993b, 2010), ofrece una nueva alternativa que conceptualiza a los instrumentos calibrados como relativos *tanto* a los datos *como* a la teoría, lo que permite acomodar diferentes perspectivas e intereses, sin poner en riesgo la necesidad pragmática de articular un programa lógico y productivo de investigación y desarrollo. Los instrumentos que miden dentro del lenguaje compartido de un marco común, hacen que los fenómenos sean reproducibles y queden disponibles para el estudio exhaustivo, al tiempo que contextualizan el valor positivo de las observaciones anómalas o de su ausencia. Los instrumentos diseñados e interpretados como objetos de demarcación encarnados, entonces, funcionan como un tipo de universal contextualizado, incipiente [inchoate] o potencial (Ricoeur, 1992, p. 289), una «relacionabilidad embebida» que funciona como «profilaxis tanto para el relativismo como para la trascendencia» (Haraway, 1996, pp. 439-440). Aunque la teoría, los instrumentos y los datos experimentales cambian a lo largo del tiempo, con ritmos variables, las comunicaciones locales enmarcan las experiencias individuales en relación con los estándares globales de forma análoga al funcionamiento del lenguaje cotidiano. Las traducciones entre la comprensión práctica de los fenómenos por parte de diferentes grupos, ocurren mediante analogías en las zonas de intercambio, ubicadas en los límites que separan a las distintas comunidades donde se forman alianzas estratégicas.

La Figura 4 (adaptada de Star & Griesemer, 1989, p. 390) entrega una visión esquemática de cómo diferentes grupos de partes interesadas se alían unos con otros en relación con propósitos compartidos a través de traducciones y objetos de demarcación. Esta visión modifica la perspectiva originalmente postulada por Callon (1985), Latour (1987, 1993a) y Law (1985). El supuesto no verificable relativo a la existencia de un único objeto abstracto y global, denominado «móvil inmutable» por Latour (1987), es dejado de lado en favor de un énfasis pragmático Quineano en los puntos de decisión observables puestos en efecto mediante una traducción al vocabulario, los conceptos y los procesos de cada grupo. Diversas partes interesadas actúan al mismo tiempo en la organización de un campo, a pesar de que los objetos de investigación especificados en cada campo habitan mundos sociales separados. Los objetos de demarcación son adaptables entre estos mundos hasta el punto que los métodos mediante los cuales son

puestos en juego vía traducción, son estandarizados dentro de los procesos de cada grupo de partes interesadas. La traducción no puede llevarse a cabo dentro de un relativismo vago y de tipo *laissez-faire,* sino que requiere el rigor de repetidas reconstrucciones del objeto común, donde cada traducción incorpora elementos de todas las demás. Los intereses de cada grupo son operacionalizados en y promovidos mediante los procesos coordinados al grado en que son incorporados transparentemente a todo el resto como participantes obligatorios en la comunidad general.

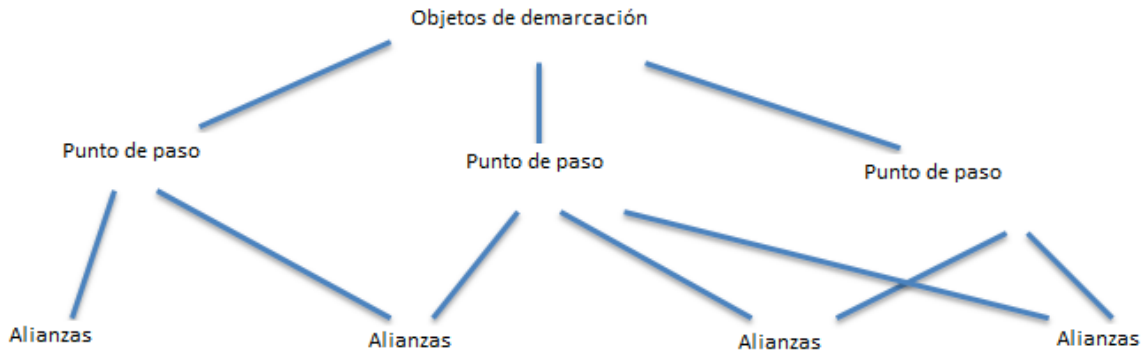

Figura 4. Alianzas y traducciones de zona de intercambio que enmarcan la inteligencia colectiva de un campo (adaptada de Star y Griesemer, 1989, p. 390).

Así, Star y Griesemer (1989, p. 390) apuntan que la coherencia de conjuntos independientes de traducciones debe ser indiferente a los procesos específicos que los producen. El número indefinido de formas en las que los actores de cada grupo de partes interesadas pueden hacer que su trabajo sea necesario para los otros grupos, da como resultado un número indeterminado de posibles traducciones. Nuestra traducción de lo que significa la *coherencia* de los objetos de demarcación dentro del dominio de la medición es efectivamente equivalente a requerir la medición de la invarianza en un modelo psicométrico. Se visibilizan nuevas posibilidades para la investigación y la práctica educacional en el contraste de los enfoques positivistas y postpositivistas con respecto a cómo se traducen los intereses de diferentes grupos entre sí.

Por ejemplo, la educación se basa en la premisa de que el entrenamiento para resolver ciertos problemas en el aula puede ser efectivo a pesar de las variaciones en la manifestación de ese ideal en estudiantes, profesores, escuelas, tests y currículos, a pesar del hecho de que ningún grupo específico de problemas podrá nunca representar todos los potenciales problemas del mundo real. Aunque las circunstancias están cambiando, la investigación y la práctica educacional contemporánea tienden a definir los objetos de demarcación y los procesos estandarizados de dos formas bastante diferentes. La manera más tradicional consiste en usar evaluaciones desarrolladas localmente y contenido curricular determinado también localmente. El paradigma es positivista, en el sentido de que los hechos objetivos (conteos y porcentajes) en cuanto a respuestas correctas e incorrectas a ciertos grupos de preguntas de evaluación se consideran suficientes para la determinación de resultados. Así, la traducción se halla cargada de problemas casi insuperables en lo referido a comparabilidad, dado que el significado de las puntuaciones cambia de modos desconocidos con cada contenido y contexto de evaluación, y entre estudiantes y currículos. Los gastos y los problemas que se enfrentarían al tratar de traducir puntuaciones de pruebas localmente definidas en aulas, grados, escuelas, distritos y regiones, superarían con creces cualquier valor que pudiera obtenerse.

El así llamado enfoque «moderno» consiste en desarrollar las así llamadas «pruebas estandarizadas» (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Plake & Wise, 2014), como el objeto de demarcación *evaluación* y en adoptar «estándares educacionales» (Troia & Olinghouse, 2013), como definición del objeto de demarcación *currículo.* De nuevo, estos instrumentos son explícitamente positivistas en cuanto a sus intenciones. Sin embargo, cada uno tiene sus propios problemas de traducción. La prueba estandarizada, como se define comúnmente siguiendo el paradigma de la teoría clásica de test, usa un conjunto común de ítems (tal vez extendido más allá de una única forma de

prueba mediante la técnica de equiparación basada en población) para establecer una interpretación con referencia a normas de las puntuaciones del test. Esto ciertamente genera una interpretación que posee un significado claro (como «el estudiante X está en el percentil *n* en la población *P*») y, además, genera estimaciones de incertidumbre para dichas aseveraciones. Sin embargo, lo que no logra hacer es entregar una interpretación clara de qué significa esto en términos del currículo mismo (como sea que este se represente), y tampoco consigue establecer la traductibilidad a otras poblaciones de partes interesadas más allá del aula o de la evaluación específica involucrada.

El uso típico de los estándares educacionales es como una lista racional que define el dominio de la prueba estandarizada asociada. Estas listas de estándares han atraído críticas (por ejemplo, véase Pellegrino, Wilson, Koenig, & Beatty, 2014; Williamson, Fitzgerald, & Stenner, 2013) de dos tipos diferentes: (a) que los estándares que se incluyen en la lista tienden a enfocarse en el conocimiento del contenido del estudiante más que de en los procesos que deben aprender para entender bien un dominio de conocimiento y (b) que son presentados como átomos individuales de este conocimiento más que como elementos enlazados en secuencias de estándares, que buscan describir el paso de los estudiantes hacia los niveles más sofisticados, los que se denominan *progresiones de aprendizaje* (o, en ocasiones, trayectorias de aprendizaje; Black et al., 2011). Ambos son serios problemas de traducción, dado que significan que la evaluación resultante no está mapeando completamente de acuerdo al currículo ni es interpretable de una manera referida a criterio por los profesores que enseñarán el currículo.

La investigación psicométrica contemporánea entrega posibilidades alternativas postpositivistas (inmodernas o amodernas) mediante un enfoque de mapeo de constructos (Wilson, 2005; Wilson & Sloane, 2000). Los modelos psicométricos probabilísticos que sustentan el mapeo de constructos operacionalizan el ideal educativo parametrizando cada faceta en el diseño general, permitiendo: (a) probar hipótesis con respecto a la representatividad de las muestras de estudiantes y de ítems, (b) teorizar sobre la coherencia de las preguntas formuladas y las respuestas recibidas y (c) caracterizar cualitativamente las medidas tanto consistentes como inconsistentes. Las evaluaciones se crearon para ser estructuradas consistentemente, lo que significa que las comparaciones son interpretables en términos comunes a través de diferentes conjuntos de ítems y de diferentes muestras de estudiantes. Los modelos matemáticos de este tipo, que articulan objetos de demarcación desde la perspectiva psicométrica, entregan nuevas oportunidades para traducir a través de los grupos de partes interesadas. En ellos, el contenido de la evaluación articula el objeto de demarcación para el desarrollador curricular, los resultados finales para el profesor y la progresión de aprendizaje para el estudiante. Dados los avances en las comunicaciones electrónica, redes, computación, algoritmos analíticos, etc., las aplicaciones creativas de los modelos psicométricos pueden conducir a importantes nuevos desarrollos en la traducción de conceptos y métodos avanzados a través los diversos grupos de partes interesadas en mejorar la calidad de los resultados educativos, lo que puede extenderse más allá del dominio de la educación, específicamente a la contabilidad para la sostenibilidad, la economía ecológica, la metrología forense y legal, entre otros campos.

El problema para todos los actores en la red de traducción es cómo reducir su incertidumbre local con respecto a cuál es el objeto de demarcación y qué significa para ellos, sin alienar a los otros grupos de partes interesadas. Permitir una traducción diferente que podría encarnar sus intereses de mejor forma, o de manera aparentemente mejor, permitirá que se convierta en un nuevo punto de paso obligatorio. Así, las traducciones ocurren mediante procesos estandarizados que coordinan diferentes perspectivas con respecto al objeto de demarcación en los puntos de paso obligatorios.

La estandarización debiera ser: menos una imposición de limitaciones arbitrarias determinadas externamente y más una forma de establecer analogías entre las perspectivas de las partes interesadas. Estas analogías deben ser capaces de replicar la extensión de los procesos de razonamiento cotidiano basado en modelos logrados en las ciencias naturales, una extensión que: (a) preserva las estructuras relacionales, (b) realiza mapeos isomorfos entre sistemas y (c) logra sistematicidad, entendida como sistemas de mapeo de estructuras relacionales de orden superior (Nersessian & Chandrasekaran, 2009, p. 186). El mapeo de constructos y el modelamiento psicométrico ayudan a organizar, revisar e implementar analogías útiles en la estandarización de los términos, para negociar puntos de paso obligatorios en redes de traducción. En efecto, el objetivo y resultado final del mapeo de constructos es la capacidad de decir que el estudiante A se relaciona con el ítem Y, del mismo modo en que el

estudiante B se relaciona con el ítem Z, digamos, con una probabilidad de éxito de 50-50. Alternativamente, la meta podría ser el decir que el estudiante A es al estudiante B como el ítem Y es al ítem Z (o como el conjunto de habilidades necesarias para tener éxito en el ítem Y es a las habilidades necesarias para tener éxito en el ítem Z).

De cualquier modo, la traducción psicométrica del objeto de demarcación como modelo matemático, la traducción analítica como datos manipulados, la traducción curricular como progresión de aprendizaje y la traducción instruccional como el qué enseñarle después a este estudiante, deben ser análogas entre sí. Los puntos de paso entre grupos de partes interesadas se vuelven obligatorios cuando encarnan el estándar aceptado para coordinar actividades. En este contexto, emerge una nueva perspectiva con respecto al valor de basar los principios de diseño de las mediciones en el teorema de la separabilidad de Rasch. Del mismo modo en que las ciencias naturales han extendido el razonamiento cotidiano basado en modelos a una combinación integrada de teoría explicativa, tests de hipótesis experimentales e instrumentos calibrados en unidades estándar (Nersessian, 2006, 2008, 2012; Nersessian & Chandrasekaran, 2009), la psicometría podría también alcanzar de mejor forma su potencial como una ciencia (en lugar de ser un conjunto de técnicas estadísticas) (Wilson, 2013a), haciendo un uso más sistemático de estas estrategias.

Las formas de abordar la medición y la calibración de los instrumentos se enfocan en la identificación y el escalamiento de constructos invariantes y unidimensionales (Rasch, 1960; Wilson, 2005; Wright, 1977) especifican las estructuras relacionales más simples, capaces de apoyar interpretaciones análogas entre grupos de partes interesadas. Esto contrasta con los modelos estadísticos que incorporan variaciones en la dificultad del ítem, dependiendo de la ubicación y consistencia de las respuestas (DeMars, 2010, p. 16), dado que estas impiden y destruyen la identificación, el mapeo y la preservación de las estructuras relacionales entre grupos de partes interesadas. De modo similar, los métodos de equiparación de tests basados en Rasch (Engelhard & Osberg, 1983; Masters, 1985; von Davier, 2010), conectan diferentes tests y evaluaciones en una red más amplia, que efectivamente constituye los mapeos isomorfos entre sistemas. Finalmente, el mapeo de constructos basado en Rasch (Wilson, 2005) facilita el diseño del ítem, la puntuación de las respuestas y el modelamiento matemático, incluyendo el modelamiento explicativo (De Boeck & Wilson, 2004, 2014; Stenner & Fisher, 2013; Stenner & Smith, 1982) que consigue la característica de sistematicidad de las estructuras relacionales de orden mayor.

La investigación tradicional de pruebas y encuestas no involucrada en el mapeo de constructos de este tipo, puntúa las respuestas sin importar que exista algún objeto de demarcación conceptualizado expresamente, suponiendo que la objetividad del procesamiento de la respuesta y la transparencia del acto de sumar puntuaciones bastan para hacer surgir los conceptos, métodos y procesos organizacionales apropiados. Luego, se imponen los estándares desde afuera, sobre la base de una autoridad social externa, y no emergen desde las experiencias de cada grupo de partes interesadas como expresiones auténticas. El valor del método de mapeo de constructos y de los modelos probabilísticos asociados, reposa sobre la capacidad de promover más efectivamente los intereses genuinos de diferentes grupos, más efectivamente de lo que podría hacerse empleando los métodos tradicionales de la teoría clásica de test y de los tests estandarizados con referencia a normas.

Es importante destacar que los grupos que participan en las redes de traducción no intentan asimilar, disolver o erradicar grupos con prioridades, culturas o idiomas diferentes. Por el contrario, necesitan multiplicidad, inestabilidad, marginalidad y multicentralidad, elementos que fomentan síntesis creativas de originalidad divergente y conformidad convergente (Berg & Timmermans, 2000, p. 38), aunque esto puede no siempre ser claro para los miembros de cada grupo. Los estudios históricos sugieren, entonces, que el éxito permanente de la ciencia se capitaliza en un equilibrio entre el pensamiento divergente y opositor, con el pensamiento convergente y unificado (Edwards, Mayernick, Batcheller, Bowker, & Borgman, 2011; Galison & Stump, 1996; Woolley & Fuchs, 2011). El énfasis positivista en la observación, el énfasis instrumentalista en la tecnología y el énfasis antipositivista en la teoría, empleados como marcos de referencia unificadores, han dado paso a la posibilidad inmoderna de una perspectiva «intercalada» que permite que los datos, los instrumentos y la teoría actúen como factores independientes pero interrelacionados, variando pesos entre y dentro de diferentes redes sociales (Ackermann, 1985; Galison, 1999; Ihde, 1991). Como observa Golinski (2012, p. 35), «Las prácticas de traducción, reproducción y metrología han tomado el lugar de la universalidad que solía suponerse como un atributo de la ciencia singular». Los investigadores de un campo específico residen típicamente

en distintas comunidades que se enfocan en asuntos específicos que involucran la instrumentación, la experimentación o la teoría. Es posible que nunca o casi nunca exista comunicación significativa entre estos grupos, más allá de la forma en que cada uno se apropia de procesos y resultados de los otros en sus propios términos (Galison, 1999).

Facilitación de la gestión de las evaluaciones por parte de los profesores

Continuando la discusión anterior, puede observarse que el evento común de un profesor aplicando una evaluación a sus estudiantes, ocurre dentro de una compleja red de objetos de demarcación y puntos de paso interconectados, además de contar con el acompañamiento de un conjunto asociado de aliados. En su mayor parte, por supuesto, todo esto ocurre sin que el profesor esté consciente de su complejidad, ni menos que invoque algo de esto. Sin embargo, aunque uno no quisiera aumentar la complejidad de la tarea de un profesor, existen buenas razones para desear que este antecedente esté disponible para el profesor y que este, al menos en parte, familiarizado con él. Por ejemplo, tener una conexión que sea lo más fuerte posible entre los resultados de evaluaciones sumativas y formativas, facilitaría el uso racional de dichos resultados en la planificación de los próximos pasos de la enseñanza. Por supuesto, esto es lo que el par mapa Wright/mapa de constructo está específicamente diseñado para lograr, y existen formas bien documentadas de la utilización de estos mapas (por ejemplo, Black et al., 2011).

La recolección de los resultados de estas evaluaciones en forma de datos accesible, empleando el mapa Wright para asignar ubicaciones a los estudiantes basadas en dicha información y ayudar a los profesores a interpretar los mapas, son todas actividades altamente complejas, y deben estar disponibles (casi) instantáneamente para ser realmente un apoyo para los profesores. Es por esto que hay otro grupo de aliados, los desarrolladores de TI, ilustrados en la Figura 4, porque sin un sistema completo de recolección y análisis de datos conectado con un sistema de interpretación guiada, un profesor de una sala de clases típica simplemente se verá superado por las demandas de ingreso de datos, manejo y análisis, todo esto sumado a la operación del sistema interpretativo. Por lo tanto, es imperativo que estos materiales e ideas se implementen en un sistema de Tecnologías de la Información diseñado para funcionar consistentemente en el marco presentado en la Figura 4.

Nuestra respuesta a esto es el sistema de evaluación formativa en línea del Centro de Investigación en Evaluación BEAR de la UC de Berkeley [UC Berkeley Evaluation and Assessment Research (BEAR)], el Software del Sistema de Evaluación BEAR [BEAR Assessment System Software] (BASS; Scalise et al., 2007; Scalise & Wilson, 2011; Torres Irribarra, Freund, Fisher, & Wilson, 2015; Wilson, Scalise, Galpern, & Lin, 2009), el cual fue explícitamente diseñado para facilitar evaluaciones autoguiadas y ha estado en desarrollo durante lo últimos 12 años. Con el apoyo económico de la Fundación Nacional de la Ciencia de EE.UU. [U.S. National Science Foundation (NSF)] y el Instituto para la Educación Científica del gobierno federal de EE.UU. [Institute for Education Science (IES)], el BASS incorpora los principios del enfoque de modelamiento de constructos descrito anteriormente, y está siendo usado actualmente en varios estados de EE.UU. para la evaluación e instrucción integrada utilizado en distintas áreas de la educación en ciencia, tecnología, ingeniería y matemáticas (STEM). El Sistema de Evaluación BEAR [BEAR Assessment System (BAS)] incluye cuatro bloques y herramientas asociadas para construir evaluaciones de calidad: Mapas de Constructo, Diseño de Ítems, Espacio de Resultados y Modelo de Medición (véase Tabla 1). Estos bloques mapean de acuerdo al Triángulo de Evaluación del Consejo Nacional de Investigación [National Research Council (NRC)], desarrollado por el comité del NRC sobre los Fundamentos de la Evaluación (National Research Council, Division of Behavioral and Social Sciences and Education, Center for Education, & Mathematics Learning Study Committee, 2001).

Tabla 2
Principios, bloques y productos del Sistema de Evaluación BEAR

Principio	Bloque	Producto
La evaluación debiera basarse en una perspectiva de desarrollo.	Constructo.	Mapa de constructo.
Coincidencia entre la práctica de la enseñanza y lo que se evalúa.	Diseño de ítems.	Ítems.
Los profesores deben ser los administradores del sistema, con las herramientas para usarlo de forma eficiente y efectiva.	Espacio de resultados.	Ejemplares y guías de puntuación.
Evidencia de calidad en términos de confiabilidad, validez y evidencia de justicia.	Modelo de medición.	Mapas Wright.

BASS es un sistema en línea para ser usado por equipos de evaluación y desarrollo curricular que trabajen con profesores de aula en el diseño, desarrollo y entrega de evaluaciones formativas basadas en progresiones de aprendizaje con sustento empírico y validación teórica (Wilson, 2004, 2009, 2013a). Al usar el software, los profesores pueden monitorear y reportar el progreso de los estudiantes de una forma que facilita el mejoramiento de dicho progreso. Los resultados se entregan en un marco que hace que los estándares de ciencia, tecnología, ingeniería y matemáticas [science, technology, engineering, and mathematics (STEM)] sean significativos y alcanzables. Una de las características distintivas de BASS es su incorporación de modelos avanzados de medición educacional para la evaluación. Este sistema accesible en línea permite a los profesores diagnosticar con precisión las necesidades de comprensión y aprendizaje de los estudiantes mediante sus módulos de evaluación, registro, análisis, retroalimentación y reportes en tiempo real.

BASS permite que profesores e investigadores modifiquen y sigan desarrollando evaluaciones mediante el uso, modificación o mejoramiento de las definiciones de constructo, los diseños de ítems, elaboración de instrumentos, la entrega de evaluaciones, la recolección de datos, la puntuación de respuestas, el modelamiento estadístico, los análisis de validez y confiabilidad y, el reporte de resultados. La Figura 5 muestra una visión general de la estructura de BASS.

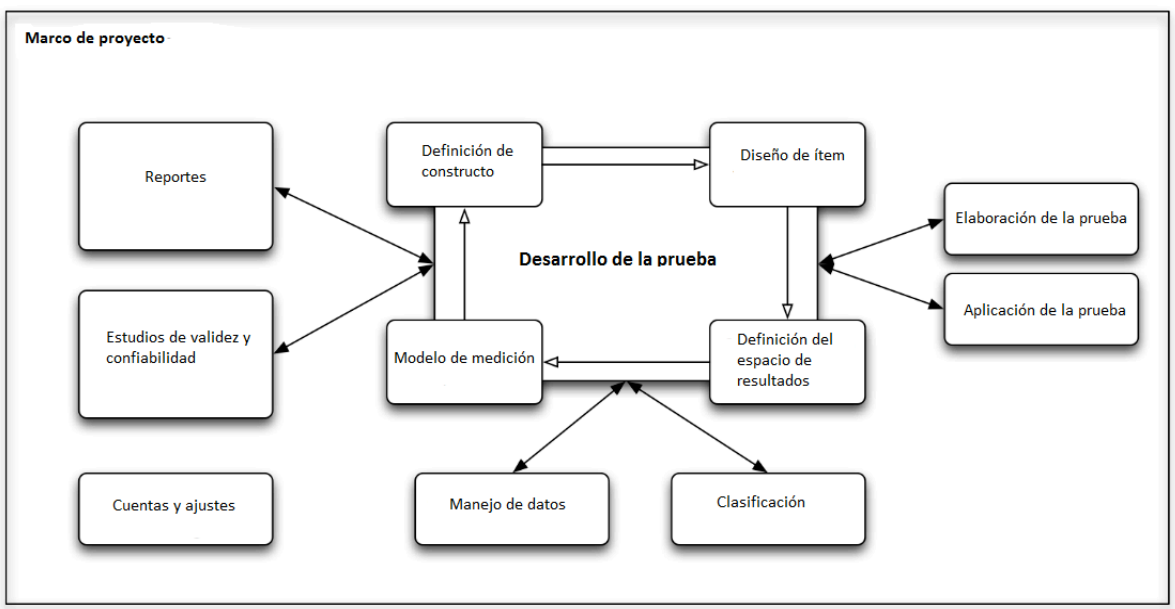

Figura 5. Visión general de los módulos del Software del Sistema de Evaluación BEAR [BEAR Assessment System Software (BASS)].

Los primeros cuatro módulos de BASS (Definición de Constructo, Diseño de Ítems, Definición del Espacio de Resultados, Modelo de Medición) despliegan la funcionalidad de los cuatro bloques fundacionales de BEAR. Los módulos quinto y sexto (Ensamblaje de Test y Entrega de Test) permiten crear y entregar instrumentos de evaluación basados en las contribuciones del ciclo de desarrollo del test. El séptimo módulo está dedicado al manejo de datos, la puntuación automática y la recodificación, mientras que el octavo (Calificación) entrega funcionalidad para todas las puntuaciones que requieren juicio humano. El noveno módulo (Estudios de Validez y Confiabilidad) incluye las características de software que se requerirán en los análisis de validez y confiabilidad. El décimo módulo (Reportes) entrega las capacidades de reporte necesarias tanto en términos de los tipos de cuadros y tablas disponibles como de los distintos tipos de datos, que en principio sería útil acceder a través de estos reportes. Finalmente, el último módulo (Cuentas y Configuración) contiene funcionalidad relacionada con la administración del sistema, incluyendo el manejo de opciones y preferencias y la asignación de permisos y acceso a los usuarios. Aunque estos módulos se hallan en distintos niveles de desarrollo y refinación, la mayoría de ellos ya están funcionales y varios han sido usados en más de un proyecto finalizado o que se encuentra en desarrollo.

Este sistema fue diseñado para entregar a los profesores un sistema para la entrega de evaluaciones formativas que accediera a los procesos cognitivos que desarrollan los estudiantes al construir respuestas a problemas que requieren pensamiento crítico. Otro propósito de BASS es promover la comprensión por parte de los profesores de los usos y técnicas de evaluación formativa y, particularmente, la integración de la evaluación formativa con el contexto interpretativo expuesto mediante mapas de constructo y su relación con estándares y progresiones de aprendizaje (Wilson, Scalise, Galpern, & Lin, 2009).

Discusión

Para que puedan ser comprendidas y útiles en formas que permitan a profesores, desarrolladores curriculares, responsables de generar políticas públicas, psicometristas y desarrolladores de evaluaciones hacer progresar el campo de la educación como un todo, las evaluaciones deben hacer mucho más que simplemente hablar a cada grupo de partes interesadas en su propio idioma. En primer lugar, deben permitirle a cada grupo verse a sí mismo y a sus intereses a través de los ojos de cada uno de los demás grupos; y, en segundo lugar, deben permitirle a cada grupo involucrarse sustantivamente con los otros haciendo que los productos de sus procesos estén disponibles para ellos de formas que ayuden a promover sus intereses. Estas traducciones son los diálogos mediante los cuales los idiomas comunes se elaboran y se orientan a materializar resultados de esfuerzos colectivos que serían inalcanzables para cada grupo por sí solo.

La psicometría, por ejemplo, debe formularse ciertos desafíos a sí misma y a sus socios de investigación para así cumplir su rol en la comunidad de reflexión. En el proyecto descrito anteriormente, los desafíos enfrentados por la psicometría podrían clasificarse como estándar y no estándar. Los desafíos estándar son aquellos que usualmente se presentan en un contexto positivista y moderno que prioriza la evidencia por sobre la teoría y el instrumento. Estos incluyen definir bien las variables, crear ítems de forma específica con respecto al diseño, desarrollar sistemas de codificación y puntuación sólidos, aplicar modelos uni y multidimensionales para interrogar apropiadamente a los datos y esfuerzos adicionales que deberían ser estándar, como confeccionar reportes útiles para los profesores, ayudándoles así a diseñar y adaptar evaluaciones al entregarles las herramientas que necesitan para ser efectivos.

Los desafíos no estándar para la psicometría incluyen incorporar la medición en modelos multidimensionales, representar y modelar enlaces entre dimensiones, desarrollando nuevos modelos que representen bien el continuo latente y/o las clases latentes en las progresiones de aprendizaje y representar y modelar el cambio longitudinal entre progresiones de aprendizaje individuales.

Los desafíos para los desarrolladores curriculares, desde el enfoque de modelamiento de constructos incluyen la necesidad de lograr una precisión conceptual mayor que la que típicamente se alcanza en la actualidad, junto con una necesidad de mayor transparencia y explicitación de las metas. Las evaluaciones de fidelidad instruccional, en general, poseen las mismas tareas, pero también pueden dotarse de herramientas para trascender los resultados meramente obvios. Asimismo, las oportunidades de desarrollo profesional pueden expandirse para los profesores en forma de nuevos modos de enlazar las prácticas de los profesores con evaluaciones a gran escala y (también) con índices de responsabilidad.

BASS se sustenta en la idea de las escalas de medición que representan estructuras generalizadas de aprendizaje. La estabilidad e inestabilidad de estas estructuras podrá comprenderse mejor en el contexto de la práctica educativa en la medida que se expandan las redes de aliados al traducir y replicar objetos de demarcación relativos a puntos de paso comunes y obligatorios. Las escalas que representan estructuras de aprendizaje estables terminarán por enlazarse en sistemas comunes, de modo similar a los sistemas metrológicos de pesos y medidas que informan al comercio y a las ciencias naturales (Fisher, 2000, 2009). Los sistemas metrológicos comunes como estos podrán, eventualmente, impactar los recursos físicos, institucionales e informativos, de formas que permitirán a educadores y empleadores trabajar mejor en conjunto para identificar y cubrir necesidades de recursos humanos y una amplia variedad de otros tipos.

Galison (1997, pp. 844-845; véase también Fisher, 2011) sugiere que la investigación debe dirigirse en este sentido. Sin hacer referencia a la formulación del problema hecha por Hayek (1948, p. 88), presentada anteriormente, Galison busca una nueva analogía capaz de informar modelos de traducción de la ciencia no unificada, de replicación y de redes de metrología. Este investigador apunta a tecnologías recientes que son más confiables y útiles en una forma algo desordenada que cuando se ordenan rígidamente (como los cristales amorfos en la electrónica y los materiales laminados en la ingeniería estructural). En este sentido, Galison evoca la metáfora de Wittgenstein sobre los conceptos como fibras entrelazadas en una hebra que es más fuerte que si estuviera formada solo por una sola fibra continua. Galison considera que esta metáfora no basta para representar los procesos dinámicos involucrados y recomienda una metáfora no mecánica que exprese la coordinación de las «diferentes

acciones simbólicas y materiales [mediante las cuales] las personas crean la cultura vinculante de la ciencia». Berg y Timmermans (2000) concuerdan, independientemente, mencionando que la estabilidad y el alcance de las redes de decisiones médicas que estudiaron «no eran producto de más instrucciones (o de instrucciones más precisas): la logística del protocolo solo podría prosperar aprovechándose parasitariamente de su propio desorden» (p. 56). Interrelaciones similares y aparentemente paradójicas de orden y desorden en el discurso (Moskowitz & Dickinson, 2002), la teoría de la interpretación (Rasch, 1992) y la percepción visual (Riani & Simonotto, 1994) sugieren que una base para una analogía productiva se puede encontrar en el fenómeno físico de la resonancia estocástica (Fisher, 1992, 2011). Caracterizada por un orden inducido por el ruido (Dykman & McClintock, 1998; Matsumoto & Tsuda, 1983; Schimansky-Geier, Freund, Neiman, & Shulgin, 1998), la resonancia estocástica presenta paralelos interesantes con el modelo de la ciencia desunificada buscado por Galison, especialmente cuando se interpreta en términos de la teoría de control no lineal (Repperger & Farris, 2010).

Heene (2013) sugiere potenciales limitaciones presentadas por la analogía de la resonancia estocástica. Heene atribuye la baja calidad de la medición en psicología al fracaso de los investigadores de exponer hipótesis del estatus cuantitativo de sus constructos a la falsificación experimental. La desunión general de la ciencia y la falta de traducciones adecuadas a través de puntos de paso obligatorios, no aparecen como factores relevantes para Heene, y estos no son mencionados en el breve comentario sobre la resonancia estocástica (Fisher, 2011) que cita. Pero de manera contraria a las aseveraciones de Heene (2013, p. 2), la analogía a la resonancia estocástica de ningún modo requiere presuponer una extrapolación de fenómenos de nivel micro a otros de nivel macro, ni tampoco esta analogía avanza principalmente como una justificación de la naturaleza probabilística de los modelos de respuesta al ítem. En lugar de aquello, como se sugiere vagamente en el breve comentario de Fisher (2011), siguiendo el método de analogía física de Maxwell (Nersessian, 2002), la idea es emplear un fenómeno físico como un punto de partida interesante para el desarrollo de teorías y de evaluación experimental. Maxwell (1965/1890, pp. 155-166, 159-160; Black, 1962, pp. 226-227; Boumans, 2005, pp. 24-25) utilizó analogías para evitar tanto la adopción prematura de una teoría explicativa, como las distracciones producto de las sutilezas analíticas a las cuales los métodos matemáticos pueden arrastrar fácilmente a los investigadores. Del mismo modo en que el modelo de un fluido sin fricción de Maxwell sirvió para hacer progresar la teoría electromagnética y su aplicación práctica, la resonancia estocástica podría apoyar el avance de la teoría y la práctica de la medición.

Heene (2013, p. 3) también declara que «actualmente existe evidencia no experimental que muestra por qué y cómo podría presentarse este error inherente del sistema en el proceso de respuesta al ítem». Esta aseveración ignora el reconocimiento de larga data de la paradoja de la atenuación (Loevinger, 1954; Masters, 1988; Sitgreaves, 1961), en la cual la remoción de la variación estocástica deriva en una estructura determinista de Guttman que no entrega información útil para estimar las distancias entre la ubicación de las personas o entre los ítems. La posibilidad de que los modelos probabilísticos Rasch se nutran de un fenómeno estructural general creando la apariencia de patrones deterministas, hace eco de pruebas de aleatoriedad irreductible, incluso en la teoría elemental de números, en la aritmética y en la física newtoniana (Chaitin, 1994). Duncan (1984, p. 220) coincidentemente observa: «Es curioso que el modelo estocástico de Rasch, que podría decirse involucra suposiciones más débiles que las que emplea Guttman [en sus modelos deterministas], realmente conduzcan a un modelo de medición más fuerte». Mientras que Guttman exige una conformidad Procrusteana con las expectativas, de modo de que todas las observaciones por debajo de la medida de una habilidad, por ejemplo, indiquen éxitos, y todas las que se encuentren por encima de la medida indiquen fracasos, las suposiciones «más débiles» de Rasch permiten una cierta variabilidad en las observaciones. Los éxitos o fracasos levemente inesperados (en el rango de probabilidades 60/40 o 40/60) no contradicen el patrón general. E incluso las anomalías marcadamente inesperadas pueden ser instruccionalmente útiles como guías cualitativas para abordar necesidades o fortalezas especiales.

Un punto mayor es que ningún modelo es verdadero, y los datos nunca se ajustan perfectamente a ellos (Box, 1979, p. 202; Rasch, 1980/1960, pp. 37-38, 2011/1973). Ningún triángulo real cumplirá jamás el teorema de Pitágoras, así como no existen péndulos matemáticos que incluyan puntos pesados suspendidos de cuerdas ingrávidas en el vacío. Como lo expresara Butterfield (1957, p. 17), «...en la vida real no tenemos bolas perfectamente esféricas que se muevan en planos horizontales uniformes; lo que pasa es que a Galileo se le ocurrió imaginarse estas cosas». En este contexto, es interesante que

Heene tome la negativa a emplear la falsificabilidad experimental en la investigación en medición como la explicación primaria de por qué la cuantificación suele ser de tan baja calidad en psicología. Es probable que Heene, como Michell (2004), suponga que la cantidad es algo que existe por sí mismo y en el mundo real, ontológicamente antes de su descubrimiento. Esto parecería negar el hecho histórico de que las cantidades medidas entran al lenguaje y al uso social mediante procesos de desarrollo. Si esa negativa se toma como la norma, entonces los modelos y las leyes no son entendidas como ideales no realistas solo aproximados por mediciones, y la falsificación exige la demostración de un estatus cuantitativo que pueden ser inferido como temporalmente anterior al marco experimental dentro del cual se manifiesta. Aunque el empirismo estricto podría satisfacerse así, ante la ausencia de una perspectiva histórica y de desarrollo, no nos queda más que una visión perfecta del pasado: el éxito en la lucha por encontrar una teoría explicativa y una unidad aditiva solo puede significar que el constructo siempre había sido cuantitativo.

Desde nuestro punto de vista, esta priorización moderna de los datos por sobre la emergencia histórica y el desarrollo de la teoría y la instrumentación constituye una falla fatal. Como ya expresaran Duhem, Quine, Feyerabend y otros, la falsificabilidad por sí misma no basta para justificar las explicaciones teóricas. Una variedad de intereses económicos, morales, políticos y de otros tipos compiten para aumentar o contradecir el peso de la evidencia relativa a sostener o cambiar las creencias de la gente sobre las cosas. Si Copérnico, Galileo y Einstein hubiesen sido empiristas estrictos, convencidos solo por la falsificación de sus hipótesis, es poco probable que hoy en día hubiésemos conocido sus nombres porque no habrían seguido persistiendo en creer en sus teorías frente a evidencia contraria. Debido a las maneras en las que los innovadores científicos son capaces de aliarse con otros en redes que promueva los intereses de muchas partes interesadas diversas (véase Latour, 1993a, en relación con Pasteur, por ejemplo), parece probable que los factores sociales que involucran zonas de intercambio, redes de traducción, alianzas de partes interesadas y objetos de demarcación jueguen roles significativos en fomentar la causa de mejorar la medición en psicología. Por otra parte, si el énfasis modernista en los datos y la falsificación llega, de alguna forma aún desconocida para ser cierto, la preocupación por estos factores sociales perderá todo sentido.

Finalmente, otra potencial limitación relacionada, que podría minar el uso de mapas de constructo como objetos de demarcación, emerge cuando Heene (2013, p. 4) declara lo siguiente: «No es necesariamente incorrecto desarrollar modelos matemáticos independientemente de las observaciones empíricas. Pero tampoco es para nada evidente que dichos modelos lograrán producir visiones empíricas». En contraposición con el supuesto tácito de Heene, tampoco es un hecho que vayan a surgir visiones empíricas de los modelos matemáticos desarrollados sobre la base de observaciones empíricas. La independencia general de los modelos matemáticos con respecto a las observaciones empíricas en la historia de la ciencia ha sido un problema filosófico reconocido por décadas. «Russell habla de casos 'donde las premisas de la ciencia resultan ser un conjunto de presuposiciones tanto empírica como lógicamente innecesarias,' y en un pasaje notable, Karl R. Popper confiesa muy claramente la imposibilidad de hacer ciencia solo basándose en elementos estrictamente verificables y justificables» (Holton, 1988, p. 41). Al contrario de las suposiciones de Heene con respecto a la primacía de los datos, Butterfield (1957) apunta que:

> La ley de inercia no es la clase de cosa que uno descubriría únicamente con métodos fotográficos de observación... requirió una forma de pensar distinta, una transposición en la mente del propio científico, ya que en realidad no vemos a los objetos ordinarios continuar su movimiento rectilíneo en ese tipo de espacio vacío (pp. 16-17).

En lugar de eso, lo que ha ocurrido repetidamente en la historia de la ciencia es que analogías que involucran proyecciones de conceptos y funciones geométricos han tenido un notable éxito en encontrar tracción empírica, accesibilidad cognitiva y encuentran un lugar en el ordenamiento socioeconómico. Esto solo refleja, en un grado significativo, el punto de Kant (1965, pp. 20-21): no simplemente seguir «las cuerdas de la naturaleza que nos guían» sino que reconocer que «la razón solo puede iluminar lo que produce siguiendo un plan propio». Aunque aquí aparecen otros problemas que requieren mucha atención (Fisher, 2003), la apropiación por parte de Rasch del método de analogía de Maxwell (Fisher, 2010) parece abrir un campo similar de posibilidades geométricas en la psicometría (Fisher & Stenner, 2013a).

La posible relevancia de la resonancia estocástica como una analogía útil también es apoyada por estudios sobre inteligencia colectiva, que indican que los campos exitosos y productivos de la investigación y de la práctica incorporan actividades divergentes de apertura y formación de enlaces en sus actividades de organización, junto con actividades convergentes de definición y delimitación, fundamentando el continuo completo en la práctica reflexiva (Woolley & Fuchs, 2011). Por lo tanto, más allá de los participantes reflexivos (Schon, 1983), se requiere de comunidades reflexivas para coordinar y alinear diversos intereses en formas que capitalicen las desuniones sistemáticas (Berg & Timmermans, 2000; Fischer, Giaccardi, Eden, Sugimoto, & Ye, 2005; Haraway, 1996), donde los miembros de diferentes grupos de partes interesadas tienen solo una comprensión parcial de los datos, de la teoría y de los instrumentos que los ayudan a potenciar sus intereses.

Los mapas de constructo, los mapas Wright y el método de modelamiento de constructos, en general, pueden entregar imágenes capaces de mediar entre las diversas necesidades de las redes de traducción. Para poder realizar esta función de manera efectiva, necesitarán encarnar analogías a nivel del sistema entre las diferentes perspectivas de los grupos de partes interesadas. Hasta la fecha, los intentos de persuadir y educar a los investigadores y al público sobre el valor de los enfoques de medición, rigurosos y significativos, no han logrado siquiera acercarse a la revolución que podría haberse esperado (Cliff, 1992), lo que ha llevado a algunos autores a sostener que una revolución como esa no puede ocurrir (Trendler, 2009). Uno de los supuestos ampliamente aceptados que apoyan esta postura se refiere a una pretendida incapacidad de realizar manipulaciones causales, lo que es contradicho por resultados de estudios realizados durante un largo período (De Boeck & Wilson, 2004, 2014; Embretson, 2010; Fischer, 1973, 1983; Stenner, Fisher, Stone, & Burdick, 2013; Stenner & Smith, 1982). Aun así, probablemente siga siendo cierto que la aplicación exitosa de modelos explicativos y predictivos también es insuficiente por sí sola para completar la tarea mayor de implementar cambios a gran escala, tendientes a mejorar la calidad de la medición en psicología. El éxito en generalizar el logro de una medición de alta calidad y la comunicación de cantidades significativas requerirá esfuerzos que vayan más allá de las pruebas y la evidencia y que implican alianzas sociales y económicas complejamente elaboradas entre grupos de partes interesadas, cuyas actividades deben coordinarse y alinearse en relación con puntos de paso obligatorios dentro de una red de traducción sistemáticamente construida.

A la manera de la instrumentación metrológicamente trazable, las unidades de medición cuantitativa y sus conceptos y vocabularios cualitativos asociados, requerirán no solo evidencia empírica y explicaciones teóricas, sino que también alianzas sistemáticas capaces de promover los intereses de cada grupo de partes interesadas más efectivamente que dichos grupos que puedan promover sus intereses por ellos mismos. El trabajo con este objetivo está avanzando (por ejemplo, Fisher & Stenner, 2013b; Mari & Wilson, 2013; Pendrill & Fisher, 2015; Wilson, 2013b; Wilson, Mari, Maul, & Torres Irribarra, 2015). Pero en lugar de ser el cumplimiento de los sueños positivistas sobre universales abstractos, o de la pesadilla antipositivista sobre dependencias locales inconmensurables, los estándares incorporados en las redes de traducción para la psicología y la educación probablemente puedan describirse con mayor realismo como una muy exigente receta de trabajo duro, la cual promete resultados productivos mayores de lo concebible hasta ahora.

El artículo original fue recibido el 18 de marzo de 2015
El artículo revisado fue recibido el 20 de agosto de 2015
El artículo fue aceptado el 21 de agosto de 2015

Referencias

Ackermann, J. R. (1985). *Data, instruments, and theory: A dialectical approach to understanding science*. Princeton, Nueva Jersey: Princeton University Press.

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

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554. doi: 10.1177/1745691612459060

Berg, M., & Timmermans, S. (2000). Order and their others: On the constitution of universalities in medical work. *Configurations, 8*(1), 31-61.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *American Statistical Association Journal, 33*(201-204), 526-536.

Best, W. R. (2008). A Rasch model of the Crohn's Disease Activity Index (CDAI): Equivalent levels of ranked attribute and continuous variable scales. En J. N. Cadwallader (Ed.), *Crohn's disease: Etiology, pathogenesis and interventions* (Chapter 5). Nueva York: Nova Science Publishers, Inc.

Black, M. (1962). *Models and metaphors*. Ithaca, Nueva York: Cornell University Press.

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research & Perspectives, 9,* 1-52. doi: 10.1080/15366367.2011.591654

Bolles, R. D. (1962). The differences between statistical hypotheses and scientific hypotheses. *Psychological Reports, 11*, 639-645.

Bond, T., & Fox, C. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences.* (3rd ed.). Nueva York: Routledge.

Boumans, M. (2005). *How economists model the world into numbers*. Nueva York: Routledge.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. En R. L. Launer, & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201-235). Nueva York: Academic Press, Inc.

Butterfield, H. (1957). *The origins of modern science* (Rev. ed.). Nueva York: The Free Press.

Callon, M. (1985). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. En J. Law (Ed.), *Power, action and belief: Sociological Review Monograph No. 32* (pp. 196-230). Londres: Routledge & Kegan Paul.

Chaitin, G. J. (1994). Randomness and complexity in pure mathematics. *International Journal of Bifurcation and Chaos, 4*(1), 3-15.

Claesgens, J., Scalise, K., Wilson, M., & Stacy, A. (2009). Mapping student understanding in chemistry: The perspectives of chemists. *Science Education, 93*(1), 56-85. doi: 10.1002/sce.20292

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3,* 186-190.

Coats, W. (1970). A case against the normal use of inferential statistical models in educational research. *Educational Researcher, 3*, 6-7.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist, 49*, 997-1003.

Daston, L. (Ed.). (2004). *Things that talk: Object lessons from art and science.* Nueva York: Zone Books.

Daston, L., & Galison, P. (1992). The image of objectivity. *Representations, 40,* 81-128.

Daston, L., & Galison, P. (2007). *Objectivity.* Cambridge, MA: MIT Press.

Dawson, T. L., Xie, Y., & Wilson, M. (2003). Domain-general and domain-specific developmental assessments: Do they measure the same thing? *Cognitive Development*, *18*(1), 61-78.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Nueva York: Springer-Verlag.

De Boeck, P., & Wilson, M. (2014). Multidimensional explanatory item response models. En S. P. Reise, & D. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 252-271). Nueva York: Routledge.

Dear, P., Hacking, I., Jones, M. L., Daston, L., & Galison, P. (2012). Objectivity in historical perspective. *Metascience, 21*(1), 11-39. doi: 10.1007/s11016-011-9597-2

DeMars, C. (2010). *Item response theory (N. Beretvas, Series Ed.). Series in Understanding Statistics*. Nueva York: Oxford University Press.

Dewey, J. (2012). *Unmodern philosophy and modern philosophy* (P. Deen, Ed.). Carbondale, Illinois: Southern Illinois University Press.

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical.* Nueva York: Russell Sage Foundation.

Dykman, M. I., & McClintock, P. V. E. (1998). What can stochastic resonance do? *Nature, 391*(6665), 344.
Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science, 41*(5), 667-690. doi: 10.1177/0306312711413314
Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches.* Washington, DC: American Psychological Association.
Engelhard, G., Jr., & Osberg, D. (1983). Constructing a test network with the Rasch measurement model. *Applied Psychological Measurement, 7*(3), 283-294.
Ewert, T., Allen, D. D., Wilson, M., Üstün, B., & Stucki, G. (2010). Validation of the International Classification of Functioning Disability and Health framework using multidimensional item response modeling. *Disability and Rehabilitation*, *32*(17), 1397-1405. doi: 10.3109/09638281003611037
Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.
Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*(1), 3-26.
Fischer, G. H., Giaccardi, E., Eden, H., Sugimoto, M., & Ye, Y. (2005). Beyond binary choices: Integrating individual and social creativity. *International Journal of Human-Computer Studies, 63,* 482-512.
Fisher, W. P. Jr. (1992). Stochastic resonance and Rasch measurement. *Rasch Measurement Transactions, 5*(4), 186-187.
Fisher, W. P. Jr. (2000). Objectivity in psychosocial measurement: What, why, how. *Journal of Outcome Measurement, 4*(2), 527-563.
Fisher, W. P. Jr. (2003). Mathematics, measurement, metaphor, metaphysics: Parts I & II. *Theory & Psychology, 13*(6), 753-828.
Fisher, W. P. Jr. (2005). Daredevil barnstorming to the tipping point: New aspirations for the human sciences. *Journal of Applied Measurement, 6*(3), 173-179.
Fisher, W. P. Jr. (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement, 42*(9), 1278-1287. doi:10.1016/j.measurement.2009.03.014
Fisher, W. P. Jr. (2010). The standard model in the history of the natural sciences, econometrics, and the social sciences. *Journal of Physics: Conference Series, 238*(012016), 1-5. doi: 10.1088/1742-6596/238/1/012016
Fisher, W. P. Jr. (2011). Stochastic and historical resonances of the unit in physics and psychometrics. *Measurement: Interdisciplinary Research & Perspectives, 9*(1), 46-50. doi: 10.1080/15366367.2011.558789
Fisher, W. P. Jr., & Stenner, A. J. (2011). Integrating qualitative and quantitative research approaches via the phenomenological method. *International Journal of Multiple Research Approaches, 5*(1), 89-103.
Fisher, W. P. Jr., & Stenner, A. J. (2013a). On the potential for improved measurement in the human and social sciences. En Q. Zhang, & H. Yang (Eds.), *Pacific Rim Objective Measurement Symposium 2012 Conference Proceedings* (pp. 1-11). Berlin, Alemania: Springer-Verlag.
Fisher, W. P. Jr., & Stenner, A. J. (2013b). Overcoming the invisibility of metrology: A reading measurement network for education and the social sciences. *Journal of Physics: Conference Series, 459*(012024), 1-6. doi: 10.1088/1742-6596/459/1/012024
Galison, P. (1997). *Image and logic: A material culture of microphysics.* Chicago: University of Chicago Press.
Galison, P. (1999). Trading zone: Coordinating action and belief. En M. Biagioli (Ed.), *The science studies reader* (pp. 137-160). Nueva York: Routledge.
Galison, P. (2008). Image of self. En L. Daston (Ed.), *Things that talk: Object lessons from art and science* (pp. 256-294). Nueva York: Zone Books.
Galison, P., & Stump, D. J. (1996). *The disunity of science: Boundaries, contexts, and power.* Palo Alto, California: Stanford University Press.
Golinski, J. (2012). Is it time to forget science? Reflections on singular science and its history. *Osiris, 27*(1), 19-36.
Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1*, 3-10.
Haraway, D. J. (1996). Modest witness: Feminist diffractions in science studies. En P. Galison, & D. J. Stump (Eds.), *The disunity of science: Boundaries, contexts, and power* (pp. 428-441). Stanford, California: Stanford University Press.
Hayek, F. A. (1948). *Individualism and economic order.* Chicago: University of Chicago Press.

Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Frontiers in Psychology, 4*(246). doi: 10.3389/fpsyg.2013.00246
Holton, G. (1988). *Thematic origins of scientific thought: Kepler to Einstein* [Revised ed.]. Cambridge, Massachusetts: Harvard University Press.
Hutchins, E. (1995). *Cognition in the wild.* Cambridge, Massachusetts: MIT Press.
Hutchins, E. (2014). The cultural ecosystem of human cognition. *Philosophical Psychology, 27*(1), 34-49. doi: 10.1080/09515089.2013.830548
Ihde, D. (1991). *Instrumental realism: The interface between philosophy of science and philosophy of technology*. Bloomington, Indiana: Indiana University Press.
Ihde, D. (1998). *Expanding hermeneutics: Visualism in science*. Evanston, Illinois: Northwestern University Press.
Ihde, D. (2012). *Experimental phenomenology: Multistabilities.* (2nd ed.). Albany, Nueva York: SUNY Press.
Kaiser, F. G., & Wilson, M. (2000). Assessing people's general ecological behavior: A cross-cultural measure. *Journal of Applied Social Psychology, 30*, 952-978.
Kant, I. (1965). *Critique of pure reason*. Nueva York: St. Martin's Press.
Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago, Illinois: University of Chicago Press.
Latour, B. (1987). *Science in action: How to follow scientists and engineers through society.* Nueva York: Harvard University Press.
Latour, B. (1990). Postmodern? No, simply amodern: Steps towards an anthropology of science. *Studies in History and Philosophy of Science, 21*(1), 145-71.
Latour, B. (1991). The impact of science studies on political philosophy. *Science, Technology, & Human Values, 16*(1), 3-19.
Latour, B. (1993a). *The Pasteurization of France.* Cambridge, Massachusetts: Harvard University Press.
Latour, B. (1993b). *We have never been modern.* Cambridge, Massachusetts: Harvard University Press.
Latour, B. (2010). A compositionist manifesto. *New Literary History, 41,* 471-490.
Law, J. (Ed.). (1985). *Sociological review monograph. Vol. 32: Power, action and belief.* Londres: Routledge & Kegan Paul.
Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*, 493-504.
Mari, L., & Wilson, M. (2013). A gentle introduction to Rasch measurement models for metrologists. *Journal of Physics Conference Series, 459*(1). doi: 10.1088/1742-6596/459/1/012002
Masters, G. N. (1985). Common-person equating with the Rasch model. *Applied Psychological Measurement, 9*(1), 73-82.
Masters, G. N. (1988). Item discrimination: when more is worse. *Journal of Educational Measurement, 25*(1), 15-29.
Matsumoto, K., & Tsuda, I. (1983). Noise-induced order. *Journal of Statistical Physics, 31*(1), 87-106.
Maxwell, J. C. (1965/1890). *The scientific papers of James Clerk Maxwell* (W. D. Niven, Ed.). Nueva York: Dover Publications.
Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept.* Cambridge: Cambridge University Press.
Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to Borsboom and Mellenbergh. *Theory & Psychology, 14*(1), 121-129. doi: 10.1177/0959354304040201
Miller, P., & O'Leary, T. (2007). Mediating instruments and making markets: Capital budgeting, science and the economy. *Accounting, Organizations, and Society, 32*(7-8), 701-734.
Moskowitz, M. T., & Dickinson, B. W. (2002). Stochastic resonance in speech recognition: Differentiating between /b/ and /v/. *Proceedings of the IEEE International Symposium on Circuits and Systems, 3*, 855-858.
National Research Council, Division of Behavioral and Social Sciences and Education, Center for Education, & Mathematics Learning Study Committee (2001). *Adding it up: Helping children learn mathematics* (J. Kilpatrick, J. Swafford, & B. Findell, Eds.). Washington, DC: National Academy Press.
Nersessian, N. J. (1996). Child's play. *Philosophy of Science, 63,* 542-546.
Nersessian, N. J. (2002). Maxwell and «the method of physical analogy»: Model-based reasoning, generic abstraction, and conceptual change. En D. Malament (Ed.), *Reading natural philosophy: Essays in the history and philosophy of science and mathematics* (pp. 129-166). Lasalle, Illinois: Open Court.
Nersessian, N. J. (2006). Model-based reasoning in distributed cognitive systems. *Philosophy of Science, 73,* 699-709.
Nersessian, N. J. (2008). *Creating scientific concepts.* Cambridge, Massachusetts: MIT Press.

Nersessian, N. J. (2012). Engineering concepts: The interplay between concept formation and modeling practices in bioengineering sciences. *Mind, Culture, and Activity, 19,* 222-239.
Nersessian, N. J., & Chandrasekaran, S. (2009). Hybrid analogies in conceptual innovation in science. *Cognitive Systems Research, 10,* 178-188.
Pellegrino, J. W., Wilson, M., Koenig, J. A., & Beatty, A. S. (Eds). (2014). *Developing assessments for the next generation science standards.* Report of the Committee on Developing Assessments of Science Proficiency in K-12. Washington DC: National Academies Press.
Pendrill, L., & Fisher, W. P. Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement, 71*, 46-55.
Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the Revised AERA, APA, NCME Standards for Educational and Psychological Testing? *Educational Measurement: Issues and Practice, 33*(4), 4-12.
Rasch, G. (1980/1960). *Probabilistic models for some intelligence and attainment tests* [Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980]. Copenhagen, Dinamarca: Danmarks Paedogogiske Institut.
Rasch, G. (2011/1973). All statistical models are wrong! Comments on a paper presented by Per Martin-Löf, at the Conference on Foundational Questions in Statistical Inference, Aarhus, Denmark, May 7-12, 1973. *Rasch Measurement Transactions, 24*(4), 1309.
Rasch, W. (1992). Injecting noise into the system: Hermeneutics and the necessity of misunderstanding. *SubStance, 21*(1), 61-76.
Repperger, D. W., & Farris, K. A. (2010). Stochastic resonance —a nonlinear control theory interpretation. *International Journal of Systems Science, 41*(7), 897-907.
Riani, M., & Simonotto, E. (1994). Stochastic resonance in the perceptual interpretation of ambiguous figures: A neural network model. *Physical Review Letters, 72*(19), 3120-3123.
Ricoeur, P. (1992). *Oneself as another*. Chicago, Illinois: University of Chicago Press.
Roberts, F. S. (1994). Limitations on conclusions using scales of measurement. En A. Barnett, S. Pollock, & M. Rothkopf (Eds.), *Operations research and the public sector* (pp. 621-671). Amsterdam, Países Bajos: Elsevier.
Scalise, K., Bernbaum, D. J., Timms, M., Harrell, S. V., Burmester, K., Kennedy, C. A., & Wilson, M. (2007). Adaptive technology for e-learning: Principles and case studies of an emerging field. *Journal of the American Society for Information Science and Technology, 58*(14), 2295-2309.
Scalise, K., & Wilson, M. (2011). The nature of assessment systems to support effective use of evidence through technology. *E-Learning and Digital Media, 8,* 121-132.
Schimansky-Geier, L., Freund, J. A., Neiman, A. B., & Shulgin, B. (1998). Noise induced order: Stochastic resonance. *International Journal of Bifurcation and Chaos, 8*(5), 869-879.
Schön, D. A. (1983). *The reflective practitioner: How professionals think in action.* Nueva York: Basic Books.
Sitgreaves, R. (1961). A statistical formulation of the attenuation paradox in test theory. En H. Solomon (Ed.), *Studies in item analysis and prediction* (p. 17-28). Stanford, CA: Stanford University Press.
Smith, E. V. Jr. (2005). Representing treatment effects with variable maps. En N. Bezruczko (Ed.), *Rasch measurement in health sciences* (pp. 247-259). Maple Grove, MN: JAM Press.
Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations,' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science, 19,* 387-420.
Stenner, A. J., Fisher, W. P. Jr., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology: Quantitative Psychology and Measurement, 4*(536), 1-14. doi: 10.3389/fpsyg.2013.00536
Stenner, A. J., & Fisher, W. P. Jr. (2013). Metrological traceability in the social sciences: A model from reading measurement. *Journal of Physics: Conference Series, 459*(012025). Recuperado de http://iopscience.iop.org/1742-6596/459/1/012025.
Stenner, A. J., & Smith, M. (1982). Testing construct theories. *Perceptual and Motor Skills, 55,* 415-426.
Taagepera, R. (2008). *Making social sciences more scientific: The need for predictive models*. Nueva York: Oxford University Press.
Torres Irribara, D., Freund, R., Fisher, W. P. Jr., & Wilson, M. (2015). Metrological traceability in education: A practical online system for measuring and managing middle school mathematics instruction. *Journal of Physics Conference Series, 588*(012042). doi:10.1088/1742-6596/588/1/012042
Toulmin, S. E. (1961). *Foresight and understanding: An enquiry into the aims of science*. Londres, Inglaterra: Hutchinson.

Toulmin, S. E. (1982). The construal of reality: Criticism in modern and postmodern science. *Critical Inquiry, 9*, 93-111.
Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology, 19,* 579-599.
Troia, G. A., & Olinghouse, N. G. (2013). The Common Core State Standards and evidence-based educational practices: The case of writing. *School Psychology Review, 42*(3), 343-357.
von Davier, A. (Ed.). (2010). *Statistical models for test equating, scaling, and linking*. (Statistics for Social and Behavioral Sciences). Nueva York: Springer.
Whitehead, A. N. (1911). *An introduction to mathematics*. Nueva York: Henry Holt and Co.
Williamson, G. L., Fitzgerald, J., & Stenner, A. J. (2013). The common core state standards' quantitative text complexity trajectory: Figuring out how much complexity is enough. *Educational Researcher, 42*(2), 59-69.
Wilson, M. (Ed.). (2004). *National society for the study of education yearbooks. Vol. 103, Part II: Towards coherence between classroom assessment and accountability*. Chicago, Illinois: University of Chicago Press.
Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, Nueva Jersey: Lawrence Erlbaum Associates.
Wilson, M. R. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*, 716-730.
Wilson, M. R. (2013a). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika, 78*(2), 211-236. doi: 10.1007/s11336-013-9327-3
Wilson, M. R. (2013b). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement, 46,* 3766-3774. doi:10.1016/j.measurement.2013.04.005
Wilson, M., Allen, D. D., & Li, J. C. (2006). Improving measurement in behavioral sciences using item response modeling: introducing item response modeling. *Health Education Research, 21*(Supplement 1), 4-18.
Wilson, M., Mari, L., Maul, A., & Torres Irribarra, D. (2015). A comparison of measurement concepts across physical science and social science domains: Instrument design, calibration, and measurement. *Journal of Physics Conference Series, 588*(012034). doi:10.1088/1742-6596/588/1/01203
Wilson, M., Scalise, K., Galpern, A., & Lin, Y.-H. (2009). *A guide to the Formative Assessment Delivery System (FADS).* Berkeley: University of California Berkeley, Berkeley Evaluation & Assessment Research Center.
Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181-208.
Wilson, T. P. (1971). Critique of ordinal variables. *Social Forces, 49*, 432-444.
Woolley, A. W., & Fuchs, E. (2011). Collective intelligence in the organization of science. *Organization Science, 22*(5), 1359-1367.
Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97-116.
Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.