

ARTÍCULO METODOLÓGICO

¿Construir o Adaptar Tests Psicológicos? Diferentes Respuestas a una Cuestión Controvertida.

Alberto Fernández, Edgardo Pérez, Ana María Alderete, María Cristina Richaud & Mercedes Fernández Liporace

1. Introducción

Durante el Congreso Argentino de Psicología del 2008, en la ciudad de Córdoba, se organizó un Simposio en el cual cuatro expertos (dos de Córdoba, dos de Buenos Aires) debatieron sobre la conveniencia de construir tests nuevos o adaptar los ya existentes. Nos pareció importante rescatar estas ponencias y ofrecérselas al lector de esta revista puesto que cada una de ellas expresa un posicionamiento diferente frente a esta problemática de notable vigencia en nuestro país.

Podría decirse, de un modo muy simplificado, que el Profesor Pérez manifiesta su preferencia por la adaptación de tests, las Profesoras Richaud y Alderete expresan una postura más favorable a la construcción de nuevos instrumentos y la Profesora Fernández Liporace opta por una posición intermedia.

A continuación, se presentan las cuatro ponencias expuestas en el congreso y, finalmente, el Profesor Alberto Fernández suministra sus propias conclusiones respecto a la problemática en discusión.

1.1. Edgardo Pérez

El problema de optar por construir o adaptar tests no posee, naturalmente, una respuesta simple y categórica. Existe una variedad de factores que pueden inclinar esa decisión crucial en una u otra dirección y, además, cada alternativa plantea desafíos considerables. En este sentido, es necesario considerar que desde una perspectiva estrictamente psicométrica ambas opciones son equivalentes en dificultad y costos.

Bandura (2001) redactó una guía para construir escalas de autoeficacia, cuyos lineamientos generales pueden extenderse a cualquier tipo de test psicológico. Estas pautas incluyen recomendaciones para la delimitación del constructo y su red de relaciones, redacción de ítems, juicio de expertos, muestreo piloto, estudios de estructura y consistencia interna, análisis de ítems, y estudios de validez de criterio. Excepto la fase teórica inicial, las demás secuencias son requeridas tanto para la construcción como para la adaptación de tests.

Probablemente la fase más compleja del proceso de construcción de un test sea justamente esta delimitación conceptual del constructo a medir. Esto requiere contar con una teoría validada del constructo y su red de relaciones causales, así como elaborar definiciones operacionales adecuadas de aquél. Este desafío es particularmente arduo para los investigadores de países que no se destacan en la producción de conocimiento teórico.

La mayoría de los tests respaldados por investigación de sus propiedades psicométricas esenciales en muestras extensas han sido elaborados en Estados Unidos (principalmente) o en Europa (en algunos casos). Adaptar alguno de estos tests implica la problemática tarea de verificar la equivalencia de sus propiedades métricas en otras culturas, pero nos asegura una teoría subyacente bien establecida. En la actualidad, ninguna revista sería acepta publicar investigaciones relacionadas con tests que no poseen un marco teórico explícito y con apoyo empírico.

En función de mis dos décadas de experiencia en construcción (Cuestionario de Intereses Profesionales, Inventario de Autoeficacia para Inteligencias Múltiples) y adaptación (Escala de Autoeficacia para la Escritura, Escala de Autoeficacia para rendimiento en Lengua, Escala de Estructura Percibida de las Metas Áulicas) de escalas, creo que la estrategia recomendable para decidir respecto a construir o adaptar tests podría formularse aproximadamente del siguiente modo: 1) Revise exhaustiva y cuidadosamente los tests existentes que miden el constructo de su interés, 2) Si esta revisión ha sido exitosa, seleccione uno de esos tests y comuníquese al autor que usted desea adaptar su test, 3) Adapte el test y mantenga la comunicación con el autor mientras realiza ese proceso, 4) No deje de publicar sus resultados, éstos seguramente allanarán el camino a otros investigadores con intereses semejantes de evaluación, 5) Sólo en el hipotético caso que no encuentre un test de calidad para evaluar el constructo de interés, construya el mismo.

En este momento estamos construyendo, junto a mi colega Jorge Mattus, una escala para medir autoeficacia para comprensión lectora en adolescentes. Luego de revisar un buen número de escalas construidas para medir este constructo llegamos a la conclusión que era conveniente construir un nuevo instrumento, puesto que la mayoría de las escalas analizadas fueron construidas ad-hoc para alguna investigación particular y, en general, no habían sido validadas de manera formal y extensiva. No obstante, consultamos con expertos internacionales en el dominio sobre la necesidad de construir una escala. Mi respuesta provisional a la pregunta que enmarca esta mesa redonda es, por consiguiente, bastante enfática: *en la mayoría de los casos adaptar tests es la mejor opción*. Borges decía que las opiniones de las personas son poco importantes puesto que cambian rápidamente. En efecto, hace diez años pensaba que era mejor construir tests y desconozco cual será mi opinión en el futuro mediato.

1.2. María Cristina Richaud

La creciente globalización mundial que se está produciendo a todos los niveles y en todas las esferas de la actividad humana, debido a la facilidad y rapidez de las comunicaciones, potencia que los instrumentos de medida generados en determinado país rápidamente se utilicen en otro distinto. Existe además un interés creciente en los estudios interculturales (van de Vijver y Lonner, 1995). Uno de los mayores problemas a los que nos enfrentamos es que, a causa de la escasez de desarrollo de instrumentos propios los importemos de otros países más desarrollados, aunque no se adapten a nuestra problemática ni a nuestra cultura. La relevancia de la cultura para la evaluación y medición psicológica es un problema importante, especialmente cuando se trasladan técnicas e instrumentos de una sociedad a otra o cuando se trabaja con personas provenientes de diferentes estratos dentro de la misma sociedad.

Los tests no son automáticamente utilizables, deben de adaptarse/construirse cuidadosamente, en función de las diferencias interculturales entre el idioma/cultura originario y aquéllos en los que se pretende utilizar. Traducir un test no es una mera cuestión de conocer bien los dos idiomas implicados, es un proceso mucho más complejo, por lo que más que de traducción habría que hablar de adaptación/construcción (Muñiz y Hambleton, 1996). La mayoría de los constructos utilizados en Psicología son altamente dependientes de

los aspectos culturales. Por lo tanto, lo primero de lo que hay que cerciorarse es que el constructo a evaluar sea equiparable en ambas culturas. Si no se da esa equivalencia razonable carece de sentido la adaptación, y si se hace se estarán midiendo constructos distintos en cada cultura.

Matesanz (1997) afirma que: “una estructura psicológica, el contenido psicológico de una dimensión o rasgo, no puede generalizarse sin más a un universo de sujetos distinto de aquel para el que se construyó el instrumento y, menos aún, a universos pertenecientes a otros países o culturas” (p.203). Pueden cometerse graves errores conceptuales cuando se traslada automáticamente un test construido en un subgrupo cultural a otro; es necesario evitar lo que se ha llamado la falacia de la categoría (Good & Good, 1985; Kleinman, 1977). Esta falacia impone la “cosificación de una categoría nosológica desarrollada para una población cultural particular y la aplicación de esta categoría a miembros de otra cultura, sin establecer su validez para esa cultura” (Good & Good, 1985, p.10). La falacia es más evidente en la imposición en nuestra cultura de, por ejemplo, categorías clínicas de enfermedad mental, tales como las del DSM o de escalas de medición de dimensiones de enfermedad mental como el MMPI, sin considerar la pregunta de si son aplicables. Debería mencionarse que la validez cultural no está establecida, ni la falacia de categoría evitada, cuando respondientes culturalmente diferentes contestan ítems de entrevistas o cuestionarios que se supone reflejan el concepto en estudio, que a menudo es tomado como el estándar.

En lo que respecta a los instrumentos utilizados en la Argentina, la mayoría son traducidos directamente por el propio interesado, sin tener en cuenta los procedimientos desarrollados para enfrentar el problema de la equivalencia de significado entre el lenguaje original y el nuevo (Brislin, 1981). De todas maneras, el problema del lenguaje no es el único a tener en cuenta. Una buena traducción, realizada con todos los procedimientos desarrollados a tal fin, no resuelve la totalidad de los temas relevantes de la medición. Al respecto, Dohrenwend (1966), refiriéndose a la tan mentada relación entre nivel económico social y desórdenes mentales, dice que implica aspectos de sustancia y aspectos de método. Y se pregunta: ¿”Cuáles son los factores culturales y situacionales que conducen a diferentes formas de expresar síntomas psicológicos?” (p.34).

Hemos podido observar en nuestras investigaciones, coincidiendo con Rogler (1989),

que las personas pertenecientes a niveles socioeconómicos bajos evaluaron los indicadores de síntomas contenidos en las escalas como menos indeseables socialmente, que los de niveles sociales más altos. Por ello, reconocen sin problemas tener determinados comportamientos que los otros jamás admitirían tener. Así, por ejemplo, frente a un ítem en el que preguntábamos “¿Habla usted con su hijo?”, el 30% de los padres de nivel socioeconómico bajo manifestó que no hablaba nunca con su hijo, mientras que sólo el 3% de padres de nivel socioeconómico medio- profesional manifestó lo mismo. Sin embargo, a partir de otras fuentes de información como el niño mismo, o los maestros, pudimos saber que, en el caso de los padres de nivel más alto, el porcentaje de los que no hablaban nunca con sus hijos era bastante mayor que el admitido, sólo que no lo manifestaban porque desde sus valores culturales, consideraban este comportamiento como indeseable. Ahondando más en el tema nos dimos cuenta de que en realidad este ítem no era un buen indicador de interacción familiar en las personas de nivel bajo, ya que la forma de interactuar verbalmente es característica de una clase social más educada, mientras que los niveles más bajos desarrollan otras formas de interacción.

En lo que hace a la interpretación de los resultados, es importante estar alerta al aplicar teoría construida con base en resultados foráneos que llevan a conclusiones erróneas. Así, por ejemplo, al analizar la percepción que los adolescentes tienen de sus padres, utilizamos una cuidadosa adaptación del CRPBI de Schaefer (1965). Sin embargo, las categorías de clasificación de los ítems fueron tomadas del autor sin modificaciones, adoptándose en la interpretación dimensiones tales como: Intrusividad, Posesividad, Aceptación, etc. Al extraer las conclusiones de nuestro trabajo, observamos que para los adolescentes argentinos era deseable que sus padres controlaran su hora de llegada, quisieran conocer a sus amigos, o les dedicaran mucha atención, mientras que estos indicadores eran calificados por Schaefer, dentro de la cultura estadounidense, como intrusividad. Obviamente, la cultura estadounidense estimula la independencia de los hijos desde muy temprano, en tanto que la latina tiene una modalidad más protectora, siendo las familias mucho más afiliativas. De no tener en cuenta estas características, se podría haber concluido que la intrusividad es una buena modalidad de relación entre padre e hijos, cuando en realidad lo que ocurre es que los mismos indicadores o ítems operacionalizan, en el caso de Argentina, el cuidado por los hijos y en el de Estados Unidos la intrusividad. Tener en cuenta

estas diferencias culturales no reduce el poder predictivo de la teoría sino que por el contrario incrementa su capacidad de generalización.

Uno de los primeros aspectos que debería tenerse pues en cuenta, en la adaptación de las pruebas, es la posible variación transcultural de la operacionalización del constructo psicológico que se quiere medir. Por esta razón se debería cuidar que la equivalencia en el contenido de los términos traducidos sea *funcional* y no literal. Otro aspecto tiene que ver con la dependencia de las propiedades psicométricas del instrumento con respecto a la población de la que se extraen los datos (Matesanz, 1997). Esto significa que las dimensiones que subyacen a un determinado constructo, como por ejemplo los componentes o ejes del análisis factorial, no siempre son estables, sino que pueden variar de acuerdo a las características de la población en la que se estudia. Este aspecto tiene particular importancia en el caso de instrumentos contruidos con estrategias empíricas, basados en criterios psicométricos como, por ejemplo, el Inventario Revisado de Personalidad (NEO-PIR) de Costa y McCrae (1992) (Richaud, Rimoldi, 1980; Richaud, 1980, 1981; Richaud, Lemos y Oros, 2003).

En el caso de las pruebas verbales, es necesario considerar también que los términos utilizados pueden adquirir diferentes significados según el contexto socio-cultural en que se empleen. Al mismo tiempo, a estos diferentes significados (*definiens*) le pueden corresponder diferentes *definienda* (constructo a operacionalizar), dependiendo de los regionalismos, modismos, frecuencia de utilización de palabras, etc. El evaluador no debe suponer que el evaluado domina el lenguaje y comprende las expresiones que se utilizan, debe asegurarse de que esto realmente ocurra. Quien responde a la prueba debe ser capaz de entender lo que está escrito “para que el ejercicio no refleje más una competencia idiomática que cualquier cosa que sea la que pretende medir la prueba” (Cohen & Swerdlik, 2001, p.58). Este último aspecto adquiere mayor relevancia en el caso de aquellas escalas originadas en el enfoque lexicográfico, el cual supone que toda cultura ha desarrollado palabras para representar en su propio lenguaje las diferencias individuales entre sus miembros (Costa & McCrae, 1998).

Numerosos autores (Matesanz, 1997; Fernández Ballesteros, 2000) subrayan la importancia de realizar una traducción precisa de los reactivos de una prueba verbal, cuidando además de mantener el sentido psicológico del enunciado, y adaptar las expresiones al contexto en que se pretende evaluar, hecho que actualmente recibe el nombre de validez

ecológica (Cohen & Swerdlik, 2001). Muchas de las pruebas verbales disponibles en nuestro idioma han sido construidas en otros países, principalmente en Estados Unidos, y traducidas y adaptadas en España. Esto podría llevar a que, por ingenuidad o comodidad, estas versiones se utilicen inapropiadamente en nuestro medio (Richaud, Lemos y Oros, 2003).

El Inventario Revisado de Personalidad NEO-PI R (por sus siglas en inglés) de Costa y McCrae (1992) es un cuestionario de personalidad de origen estadounidense, traducido y adaptado al español, y que ha sido utilizado en nuestro país sin una previa adaptación. Este error responde a la creencia de que adaptar una prueba significa sólo traducirla idiomáticamente, perdiendo la esencia de lo que significa evaluar y medir válidamente. La versión realizada en la Argentina (Richaud, Lemos y Oros, 2003) fue enviada en el año 1996 a Costa y McCrae para que autorizaran la traducción española. Los autores informaron que ya existía una versión española del cuestionario y que, a menos que hubiera un regionalismo que lo justificara, debía utilizarse esa versión. Por este motivo, se decidió realizar una comparación ítem por ítem de la escala española ya existente con la recientemente creada. Al comparar ambas versiones, encontramos diferencias importantes en 65 de los 240 ítems.

En vista de estos hallazgos, y como es fundamental que una prueba verbal exprese exactamente los procesos evaluados de una manera inteligible para el examinado, a riesgo de perder validez constructiva, se reforzó la necesidad de realizar la adaptación argentina del instrumento. Para ello, se presentaron los 65 ítems con diferencias, en ambas versiones española y argentina, a los 100 jueces argentinos, quienes debieron compararlos de acuerdo a los siguientes criterios: equivalencia, comprensión, redacción y habitualidad (Richaud, Lemos y Oros, 2003). Se analizó la adhesión de respuestas a cada categoría, utilizando ji cuadrado para evaluar si existía una preferencia significativa hacia alguna de las opciones. Para el estudio léxico del cuestionario, se comparó la versión española ya existente con nuestra versión argentina, encontrando que cerca del 30 % de los ítems ($n= 65$) estaban expresados de manera muy diferente. Tras analizar las respuestas de los jueces se halló que 42 de los 65 ítems de la versión argentina, fueron aceptados por consensos como más inteligibles, mejor redactados y de expresión más usual, por lo que consideramos que podían reemplazar a los 42 ítems correspondientes de la traducción española original.

Se analizaron los resultados de los 23 ítems restantes, en función de los criterios de equivalencia y claridad verbal (comprensión, redacción y habitualidad). Se consideró que el

criterio de equivalencia reviste mayor importancia por ser el que afecta más directamente a la validez. Por ejemplo, el ítem: “Disfruto resolviendo adivinanzas que me hacen pensar” no significó lo mismo para los jueces que el ítem “Disfruto resolviendo rompecabezas enredados”. Ambos enunciados pretenden operacionalizar el factor *Apertura*, pero los ejemplos dados involucran procesos mentales diferentes. Así, una adivinanza implica mayormente procesos cognitivos, mientras que un rompecabezas supone además una destreza motriz. Por esta razón, se sugirió incluir ambos aspectos en un nuevo reactivo más general, “Disfruto resolviendo acertijos (adivanzas, rompecabezas) que me hacen pensar”.

Aparte de la adecuación de los estímulos o ítems, es importante el tipo de formato de los mismos y el concepto de velocidad en la ejecución de las tareas. Hay formatos, como el de elección múltiple, muy habituales en ciertas culturas, que no lo son tanto en otras. Los occidentales tendemos a asumir que no sólo hay que hacer las cosas bien, sino rápido, y así somos socializados desde niños. Pero esta asunción no se da en todas las culturas (van Leest & Bleichrodt, 1990; van de Vijver & Poortinga, 1991), y cuando se les pide que trabajen en la prueba lo más rápidamente posible, no entienden qué apuro puede haber para acabar pronto. Hay que tratar de minimizar los efectos de la velocidad de respuesta, siempre y cuando no forme parte del constructo a evaluar.

A medida que la distancia cultural aumenta, más probabilidades hay de que haya diferencias entre los constructos. Por ejemplo, una variable tan clásica como la inteligencia tiende a ser conceptualizada de muy diverso modo de unas culturas a otras; mientras en Occidente suele asociarse a eficiencia y rapidez, en algunas sociedades del Este se asocia con actitudes reflexivas y reposadas, más que con rapidez (Lonner, 1990). Es necesario asegurarse también de que los contenidos, formatos, estímulos, gráficos, diagramas, figuras, unidades de medida, y signos en general, son igualmente familiares para las culturas implicadas.

Otro aspecto clave es que los constructores recojan datos de todo tipo, que permitan evaluar la validez y comparabilidad de las dos versiones. La tecnología estadística es actualmente potente y variada para analizar los datos, pero no hay que olvidar que cuanto más amplias sean las muestras utilizadas, con más eficacia funcionan las técnicas estadísticas. Estas técnicas son complementarias de los correspondientes análisis racionales, no los

sustituyen. En los últimos años se ha perfeccionado notablemente la tecnología para evaluar el sesgo de las puntuaciones (Camilli & Shepard, 1994; Fidalgo, 1996; Holland & Wainer, 1993), pero sigue siendo de suma importancia el análisis racional por parte de un equipo de expertos en las culturas implicadas. Lo que las técnicas detectan es que cierto ítem funciona de modo diferente en un grupo que en otro, por lo que suelen agruparse bajo el nombre genérico de Funcionamiento Diferencial de los Ítems. Será el psicólogo quien tenga que analizar las razones de ese funcionamiento diferencial. En síntesis, hay que garantizar tanto por medios estadísticos como analítico-rationales que el test adaptado es igualmente válido y confiable que el original.

En los casos de estudios transculturales donde es necesario utilizar pruebas equivalentes, ayudará notablemente a la interpretación si se dispone de la documentación apropiada sobre todo el proceso de construcción y aplicación, diseños utilizados, métodos para la evaluación de la equivalencia entre las dos versiones, ítems que se han modificado y por qué, etc. Así y todo se recomienda prudencia extrema a la hora de interpretar las diferencias halladas entre poblaciones distintas. No se pueden interpretar directamente sino que hay que fundamentar esas diferencias, cuyas posibles causas pueden ser múltiples. Por ejemplo, cuando se compara el rendimiento académico de estudiantes de países desarrollados con otros de países en vías de desarrollo, pueden atribuirse simplista e irresponsablemente las diferencias a diferencias reales en las correspondientes aptitudes, cuando en realidad existen toda una serie de factores distintos, tales como las políticas educativas, la currícula, el acceso a la escolarización, equipamientos y recursos educativos disponibles, idioma materno frente al escolar, etc., por citar sólo algunos ejemplos. Es necesario tener mucha prudencia a la hora de interpretar las diferencias. La tecnología psicométrica desarrollada bajo el enfoque de la teoría de respuesta a los ítems (Hambleton et al., 1991; Muñiz, 1996b; Navas, 1996), permite establecer equiparaciones entre las puntuaciones con bastante precisión.

Adaptar pruebas psicológicas con el viejo procedimiento de traducirlas textualmente y construir normas estadísticas adecuadas a la nueva población, basándose en el viejo modelo psicométrico aditivo probabilístico, es completamente inadecuado y desaconsejable. En general, lo más pertinente sería construir nuevas pruebas que operacionalicen exactamente el concepto, atributo o procesos que el investigador desea estudiar, dentro de su contexto teórico y cultural y siguiendo los principios psicométricos adecuados.

1.3. Ana María Alderete

Las presentes reflexiones no surgen desde la visión o perspectiva de un especialista en psicometría sino desde la práctica cotidiana en la intervención profesional y principalmente en la investigación. Mi primera participación en investigación fue en un proyecto de caracterización de estudiantes universitarios para lo cual se utilizaban entre otros instrumentos, cuestionarios de intereses vocacionales elaborados en EE UU. La observación detallada de los ítems mostraba que las actividades que se presentaban distaban bastante de las actividades cotidianas de nuestros jóvenes (un análisis de contenido corroboró esto). Este tipo de dificultades las he observado a lo largo de mi trabajo como investigadora, determinando en muchos casos la necesidad de elaborar nuestros propios instrumentos.

De los sesgos (Van de Vijver y Tanzer, 1997) que se pueden producir al utilizar pruebas elaboradas en otras culturas me parece importante centrar la atención en dos aspectos:

En primer lugar, cuando en la definición conceptual del constructo que pretende evaluar el tests encontramos sesgos culturales, es fundamental tener en cuenta la importancia que cada cultura otorga a las conductas o aspectos que se encuentran incluidas en esa definición. Un error común es pensar el constructo haciendo referencia a las personas, niños, o jóvenes en abstracto como si presentaran características universales sin tener en cuenta el contexto ecológico y cultural en donde desarrollan su vida cotidiana. Esto ha sido muy señalado en áreas como la psicología clínica (lo que en nuestra cultura es considerado patológico en otras es considerado un “don especial”), en psicología política (por ej. la valoración diferente de ciertas conductas según los grupos). El otro aspecto está referido al sesgo que puede producirse por la inadecuación de los ítems tanto en lo referido al lenguaje como a las pautas culturales de los grupos a ser evaluados; en efecto, algunos ítems poseen significados diferentes en distintas culturas.

En el primer caso la adaptación de la prueba no sería posible desde lo metodológico pues es necesaria una redefinición conceptual generando de este modo un nuevo instrumento. En el segundo caso, que es el más común, la adaptación de la prueba requiere la elaboración de nuevos ítems y aquí cabe la pregunta ¿hasta dónde se trata del mismo instrumento? ¿Los nuevos ítems son los indicadores operacionales pertinentes para evaluar el constructo

originalmente definido? En muchos casos las diferencias encontradas en los estudios de la estructura interna de la prueba se deben a la inadecuación de los nuevos indicadores para evaluar el constructo. No cabe duda que lo fundamental es la adecuada definición tanto conceptual como operacional del constructo y la identificación de la red de las variables relacionadas.

En relación a la pregunta inicial, en mi experiencia en investigación, en muchos casos fue necesario elaborar nuevos instrumentos. Entre ellos, cabe señalar el Cuestionario de Sintomatología Mental, pruebas de diagnóstico del preescolar para niños de sectores carenciados, un inventario de pautas de crianza y una escala de vulnerabilidad política. También nos vimos en la necesidad de adaptar algunos instrumentos como, por ejemplo, la Encuesta de R.Valet.

Podría concluirse que no existe una única respuesta, cualquiera de los dos caminos implica un trabajo arduo a largo plazo. A favor de la adaptación de pruebas se sostiene la ventaja de contar, en la mayoría de los casos con teorías sólidas y la posibilidad de realizar estudios comparados. Es mi postura que si los dos aspectos señalados anteriormente no se presentan, el camino es adaptar una prueba que ha mostrado solidez. Pero, como generalmente hay una fuerte influencia de las variables socioculturales en la mayoría de las situaciones el camino es el de construir nuestros propios instrumentos.

1.3. Mercedes Fernández Liporace

Tanto los usuarios de tests como los investigadores nos vemos afectados por la escasez de tests debidamente adaptados a las peculiaridades de las poblaciones con las que trabajamos. Si bien es cierto que en el ámbito de la investigación resulta deseable que los instrumentos de recolección de datos se mantengan lo más inalterados que sea posible (Jensen, 1983) a fin de comparar resultados en estudios diversos con similitud de condiciones metodológicas, quienes desempeñamos habitualmente estas actividades reconocemos que respetar este tradicional precepto puede llevarnos, la mayoría de las veces, a cometer errores de extrema gravedad en cuanto a la valoración final que hagamos de los resultados obtenidos.

Teniendo en cuenta que el establecimiento de equivalencias en el constructo en estudio y, secundariamente, el problema del lenguaje –tanto verbal cuanto icónico - son aspectos a los que, indubitablemente debemos atender, las cuestiones métricas pasan a ser las

menos complicadas, aunque debe observarse que ambas se ven afectadas por las dos anteriores. Y dado que las condiciones en que este constructo se operacionalice se hallarán íntimamente ligadas al contenido de los ítems que, a su vez, determinará, en buena parte, las respuestas de los sujetos, la influencia de factores culturales (micro y macro) es un tema que no puede dejarse a un lado. Los comportamientos humanos se hallan mediatizados e, incluso, moldeados por tales variables culturales, al igual que el lenguaje en todas sus manifestaciones. Es por ello que la delimitación y el análisis de los constructos que se intenta evaluar también se hallan determinados, circularmente, por las respuestas de los sujetos a los reactivos que, más allá de la precisión de las traducciones y adecuaciones en los modismos y expresiones, será una reacción no sólo dada por determinantes psíquicos sino también culturales.

Más allá de estas especulaciones, la investigación científica se dirige, en última instancia, a la formulación de generalizaciones y leyes y, según ese espíritu, el método debería llevarnos a trabajar, en un mismo recorte temático, con instrumentos inalterados – o con mínimas alteraciones -, que den cuenta de constructos equivalentes en varios contextos y poblaciones, aunque también debe reconocerse que este estado de situación rara vez acontece. Expresiones tales como tests libres de cultura (Jensen, 1980; Rivers, 1901), tests culturalmente justos (Samuda, 1975) y tests de reducida influencia cultural (Flynn, 1984) parecen ser sólo ideas regulativas a las que aspiramos, aceptándose desde hace tiempo la imposibilidad de eliminar la influencia cultural en las pruebas psicométricas (Greenfield, 1998; Schooler, 1998; Wooper, 1975), y aún prefiriéndose tomar las distinciones fenoménicas pertinentes en un contexto dado como aspectos que enriquecen la descripción del objeto de estudio antes que como fuentes de error. Tampoco debería perderse de vista que circunscribirse únicamente a tales distinciones, pertinentes sólo en un contexto cultural específico, restringe enormemente la posibilidad de generalización de las conclusiones y ello va, por ende, en contra de los objetivos últimos de la ciencia; cada instrumento debería ser examinado en profundidad, tomándose decisiones particulares y bien fundamentadas según el caso, puesto que adaptar y construir implican prácticamente la misma cantidad y complejidad de operaciones y tareas.

Debería atenderse, asimismo, como urgencia prioritaria, a las carencias de formación

e información que los usuarios de tests en el ámbito de aplicación tienen sobre estas cuestiones. Un importante desafío para quienes, por trabajar en investigación con estos instrumentos, conocemos más a fondo estas dificultades reside en atrevernos a salir hacia la comunidad profesional para alertarla sobre adaptaciones deficitarias y prevenirla contra errores graves que, en definitiva, podrían afectar la vida de personas reales y concretas.

2. Conclusiones

2.1. Alberto Fernández

Como puede apreciarse a través de la lectura de los textos previos las posturas de los distintos psicometristas respecto de la adaptación o construcción de tests no son unívocas. Mientras algunos se inclinan más por la construcción de tests, otros abogan por la adaptación de los mismos. Quienes se inclinan por la primera posición fundamentan la misma en el argumento, sólidamente demostrado con investigaciones empíricas, de los numerosos errores en los que se incurre al adaptar, o simplemente traducir, un test sin tener en cuenta las numerosas diferencias culturales existentes entre ambas culturas. La más importante, como ya ha sido señalada, es el sesgo de constructo. Adaptar un test de una cultura a otra simplemente asumiendo que el constructo medido en la cultura de origen se presenta de forma idéntica en la cultura meta es un serio error que puede provocar una cadena de errores subsiguientes. No debemos olvidar que los tests psicológicos son nuestros instrumentos de medición y que se utilizan para realizar investigaciones en las que se ponen a prueba hipótesis psicológicas de las que se derivan teorías de las cuales finalmente se derivan intervenciones en los diversos campos de acción del psicólogo. Por otra parte, muchos de esos tests tienen aplicación práctica en diversos campos de intervención psicológica: clínica, legal, social, educacional, etcétera. De modo que basar intervenciones en tests que miden constructos que no existen o están pobremente representados en nuestra cultura generará un significativo deterioro de nuestra práctica profesional con las lógicas consecuencias para las personas que hayan sido sometidas a dichas prácticas. No es posible entonces soslayar la enorme importancia y solidez del argumento de quienes abogan por la construcción de tests antes que la adaptación.

La posición alternativa que defiende la conveniencia de adaptar tests tiene también un serio respaldo argumental, la cual es la escasez de desarrollo teórico en nuestra cultura que

pueda generar el suficiente cúmulo de investigación necesario para el desarrollo de un instrumento de medición psicológica. En este caso se aduce que la contrastación empírica de una teoría psicológica es un proceso laborioso que requiere de muchos recursos de los que no se dispone fácilmente en nuestro medio. Este argumento es también muy consistente ya que una simple inspección de la bibliografía publicada en nuestro idioma nos permitiría comprobarlo.

Como señalé anteriormente, quienes se inclinan por la construcción de tests tienen sólidos fundamentos, sin embargo, a través de esta posición se corre el riesgo de caer en un relativismo cultural. Dentro del marco del relativismo cultural no habría valores culturales universales y, por lo tanto, es necesario estudiar cada cultura como si fuese única. La práctica de “importar” teorías e instrumentos es considerada errónea. En mi opinión dicho relativismo cultural es una postura poco sustentable teóricamente- hay mucha investigación que demuestra que muchos constructos son equivalentes entre distintas culturas-, y hasta contradictoria en la práctica. Como ejemplo de esta última afirmación basta decir que la mayor parte de las teorías psicológicas vigentes, psicoanálisis, guesalt, cognitivismo, teoría sistémica, por mencionar solo algunas; han sido creadas en otras culturas. Dichas teorías fundamentan la inmensa mayoría de los trabajos científicos realizados en todo el mundo. Aun más, la teoría psicométrica que nos brinda los elementos teóricos y técnicos para poner a prueba a las teorías psicológicas tampoco es un desarrollo cultural propio. El texto que marca los lineamientos básicos que se deben seguir para utilizar y evaluar los instrumentos de medición psicológica, y por el cual se rigen la mayor parte de los psicometristas en el mundo son los “Standards for Educational and Psychological Testing” (1999). Dicho texto representa la posición teórica de tres instituciones estadounidenses: American Educational Research Association, American Psychological Association, National Council on Measurement in Education. De tal modo que sostener una posición de relativismo cultural coherente implicaría abandonar todas las teorías y prácticas psicométricas utilizadas en la actualidad para desarrollar alternativas indigenistas a las mismas. Tal posición no solo parece extremadamente laboriosa sino también fútil dado que es indiscutible el hecho de que hay principios psicológicos que son universales (Keith, 2011).

En síntesis, considerando los fundamentos de ambas posiciones creo que la opción

más conveniente parece ser la adaptación de tests siempre y cuando la misma involucre un riguroso proceso de investigación que asegure la equivalencia de constructos en primer lugar y una correcta adaptación de la escala para evitar toda fuente de sesgos. La construcción sería una alternativa cuando la adaptación no es conveniente debido a la no equivalencia transcultural de los constructos o al hecho de que las características del instrumento que se quiere adaptar haga que ese proceso sea tanto o más laborioso que crear uno nuevo. Al respecto, la Comisión Internacional de Tests (International Test Commission) ha elaborado una serie de pautas para adaptar tests de una cultura a otra (Hambleton, 2005). Dichas pautas han sido puestas a prueba con éxito (Hambleton, Yu & Slater, 1999) lo cual demuestra que esta es una opción válida en el controvertido tema de la adaptación de pruebas psicológicas.

Bibliografía de consulta resumida:

- American Educational Research Association (1999) *Standards for educational and psychological testing*. Washington, DC.
- Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. En R. K. Hambleton, P. Merenda y C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Mahwah, NJ: Erlbaum.
- Hambleton, R. K., Yu, J., & Slater, S. C. (1999). Field-test of the ITC guidelines for adapting educational and psychological tests. *European Journal of Psychological Assessment*, 15(3), 270-276.
- Keith, K. D. (2011) Introduction to Cross-Cultural Psychology. En: Keith, K.D. (Eds.), *Cross-Cultural Psychology: Contemporary Themes and Perspectives*. West Sussex: John Wiley & sons.