ISSN - 1659-2921

5

CUADERNOS METODOLÓGICOS

LA ENTREVISTA COGNITIVA:
Guía para su aplicación en la evaluación y mejoramiento de instrumentos de papel y lápiz

Vanessa Smith-Castro
Mauricio Molina Delgado

Aniversario 50 IIP - UCR

LA ENTREVISTA COGNITIVA:
GUÍA PARA SU APLICACIÓN EN LA EVALUACIÓN Y MEJORAMIENTO DE INSTRUMENTOS DE PAPEL Y LÁPIZ

Vanessa Smith-Castro; Mauricio Molina Delgado

Serie Cuadernos Metodológicos.
San José, CR.: Instituto de Investigaciones Psicológicas, Universidad de Costa Rica.

ISSN 1659-2921.

Smith, Vanessa; Molina, Mauricio
Cuaderno Metodológico 5. La entrevista cognitiva: guía para su aplicación en la evaluación y mejoramiento de instrumentos de papel y lápiz. San José, CR.: Instituto de Investigaciones Psicológicas, Universidad de Costa Rica. 2011.

CONTENIDO

INTRODUCCIÓN

Mucha de la investigación que hacemos en las Ciencias Sociales, de la Educación y de la Salud se construye a partir de información obtenida a través métodos de autoreporte. En estos métodos consultamos directamente a las personas sobre el tópico o constructo que queremos medir y una de las principales formas de hacerlo es mediante cuestionarios de papel y lápiz.

Al utilizar cuestionarios autoaplicados, los y las investigadores suponemos que los reactivos posibilitan a las personas reaccionar tal y como lo esperamos. Esto es, que las personas comprenden los ítems de manera adecuada, que activan la información apropiada para responderlos, que juzgan sus respuestas "debidamente" y que responden a los reactivos en los distintos formatos de respuesta que les ofrecemos de manera acertada. Todo esto según nuestras expectativas, proyecciones y criterios derivados de las teorías con que trabajamos.

Cualquiera que haya intentado desarrollar, validar o adaptar un instrumento de este tipo, reconocerá que estos supuestos están muy lejos de cumplirse en la realidad. Hasta los más experimentados investigadores e investigadoras se han tenido que enfrentar con el hecho de que sus ítems no fueron comprendidos tal y como ellos y ellas esperaban. Debido a estos problemas de comprensión, en la estructura de sus escalas aparecen dimensiones o factores ajenos al constructo que pretendían medir, y la presencia de esos factores de método disminuyen la confiabilidad de sus escalas y la utilidad de sus inferencias.

Lamentablemente, los y las investigadoras se enteran de todo esto una vez que han aplicado el cuestionario en el estudio piloto, después de una considerable inversión de tiempo y dinero. Así, nos vemos obligados a descartar los reactivos “malos” usando los criterios técnicos recomendados por la psicometría, pero sin tener claridad sobre la fuente de la imperfección de los reactivos, sospechando eso sí que están involucrados problemas de comprensión que pudieron ser evitados.

El objetivo del presente cuaderno metodológico es ofrecer una herramienta para analizar los mecanismos cognitivos involucrados en el proceso de contestar las preguntas de un cuestionario, y así detectar problemas en los distintos momentos de este proceso antes de pasar a estimar las propiedades psicométricas del instrumento en el estudio piloto. Esta herramienta se conoce como Entrevista Cognitiva (de ahora en adelante EC).

La EC es un nombre genérico para describir un dispositivo de evaluación del proceso de respuesta, y consiste una serie de entrevistas individuales semiestructuradas en ambiente controlado con una muestra pequeña de la población meta. Durante las entrevistas, las personas participantes completan el cuestionario en estudio y realizan una serie de pruebas para detectar problemas a la hora de contestarlo (Willis, 2005).

Este cuaderno pretende ofrecer una guía práctica para la implementación de la EC como parte de las estrategias para desarrollar, adaptar, mejorar o validar un instrumento de papel y lápiz. No pretendemos hacer un tratamiento exhaustivo o extenso de todos los aspectos que podrían tratarse alrededor de esta técnica.

Deseamos más bien ofrecer una guía accesible, sencilla y muy ilustrativa de los alcances de este método.

El texto se divide en cinco secciones. En la primera sección hacemos un repaso breve sobre conceptos clave de medición y psicometría, y resaltamos el lugar de la EC dentro de las diferentes evidencias de validez y confiabilidad que los investigadores podemos recabar a la hora de desarrollar, validar o adaptar instrumentos.

En la segunda sección nos concentramos en los principios teóricos de la EC. Aquí hacemos énfasis en algunos de los principales conocimientos que las ciencias cognitivas han amasado en los últimos años sobre la arquitectura y los mecanismos del aparato cognitivo que impactan en la forma en que las personas contestan instrumentos de papel y lápiz.

En la tercera sección nos concentramos en los procedimientos concretos de la EC. En esta sección presentamos las diferentes pruebas cognitivas que componen una EC y ofrecemos varios ejemplos de nuestra práctica cotidiana en donde hemos aplicado la EC con éxito para producir instrumentos más preciso y sensibles a lo que queremos medir. En esta sección tratamos también sobre los aspectos logísticos de la EC, es decir, los aspectos más prácticos de su aplicación, incluyendo los procedimientos para procesar la información que surge de las entrevistas cognitivas y las decisiones que podemos tomar a partir de las mismas.

Si bien esta técnica tiene una amplia aceptación, son pocos los estudios que se han dedicado a documentar el impacto de la EC sobre los instrumentos

utilizados en el campo. Es por ello que la cuarta sección la dedicamos a presentar los resultados de una prueba empírica especialmente diseñada para estimar la utilidad de la EC y su impacto en las características psicométricas de instrumentos de papel y lápiz. Esta sección nos permite mostrar empíricamente el valor de esta técnica para mejorar instrumentos de papel y lápiz.

En la quinta y última sección hacemos, a modo de cierre, un par de reflexiones sobre las ventajas de esta técnica, sus alcances y limitaciones y su papel en el proceso de desarrollo, validación o adaptación de instrumentos de papel y lápiz.

En las primeras cuatro secciones hacemos recomendaciones específicas sobre literatura complementaria que puede ser de gran utilidad para profundizar sobre los aspectos que aquí no queden claros o resulten insuficientes para el lector o la lectora.

Este texto está pensado para investigadores, estudiantes y profesionales que deseen desarrollar y/o adaptar instrumentos de papel y lápiz, ofreciéndoles una herramienta de fácil aplicación para recopilar evidencias sobre el funcionamiento de los reactivos de sus instrumentos y producir un conocimiento más fiable y preciso sobre la realidad que investigan.

La investigación que fundamenta este texto se desarrolló gracias al financiamiento otorgado por el Sistema de Fondos del Consejo Nacional de Rectores de Costa Rica durante el 2009 y el 2010. La redacción del texto fue posible gracias a los Fondos Para Pasantías de la Vicerrectoría de Investigación de la Universidad de Costa Rica, otorgados a la primera autora. Un especial

agradecimiento se merece el Dr. Domingo Campos, quien nos hizo percatarnos de la utilidad de la EC como herramienta para mejorar nuestros instrumentos de medición. Igualmente agradecemos profundamente al M.Sc. Alfonso Villalobos, quien trabajó con nosotros en las etapas iniciales del proyecto de investigación.

La idea de realizar una investigación sobre la EC y sus bondades surgió del seminario de verano "¿Qué hacemos cuando contestamos un cuestionario?: Psicología cognitiva y el diseño de instrumentos de papel y lápiz" que organizamos en febrero del 2007. Allí nos reunimos con estimados colegas y estudiantes provenientes de diversas disciplinas como la estadística, la lingüística, la psicología y la sociología, entre quienes destacamos a la Licda Marjorie Moreno, a la Licda. Laura Sánchez, al M.Sc. Olman Ramírez, al Dr. Manuel Arce y al Dr. Napoleón Tapia. A todos ellos nuestras más sinceras gracias por las amenas y apasionantes discusiones que nos motivaron a realizar este proyecto.

No podemos olvidar a los asistentes de investigación del IIP Armel Brizuela, Andrés Carvajal, Natalia Galeano, Pía Mehler, Jorge Morales, Sally Schulze y Fabiana Zúñiga, por su colaboración en todas las etapas del proyecto de investigación. Sin su apoyo esta investigación simplemente no habría podido ser posible. Armel Brizuela nos apoyó también con la revisión Filológica del presente texto. También agradecemos a los licenciados Camilo Retana, Lucy Gutiérrez, Maritza Quesada y al Dr. Rolando Pérez por su colaboración en etapas críticas del proceso de investigación.

Finalmente agradecemos a los y las colegas del IIP por su valiosa retroalimentación a las versiones preliminares de este texto, a la serie Cuadernos

Metodológicos de dicho Instituto por ofrecernos este espacio para publicar, a las autoridades universitarias que nos dieron los permisos respectivos para llevar a cabo diversas encuestas en el Campus Rodrigo Facio y en particular a todas las personas que muy amablemente tomaron parte de su tiempo para contestar nuestros cuestionarios.

ALGUNOS ASPECTOS BÁSICOS A RECORDAR SOBRE MEDICIÓN Y PSICOMETRÍA

Conceptos científicos como estrés, depresión, inteligencia, frustración, prejuicio, salud mental, emociones, actitudes, ansiedad, motivación o estrés laboral no tienen una existencia concreta similar a las entidades físicas que se ofrecen a nuestros sentidos. Son conceptos que sobrepasan la observación empírica y muchas veces expresan entidades teóricas. A tales conceptos se les llama "constructos", "conceptos no observacionales" o "variables latentes" (ver Bollen, 2002, Montero, 2008).

Nadie ha visto ni ha tocado nunca la motivación de logro de las personas, pero si puede inferirla a partir de indicadores, como su persistencia, para terminar una tarea. Es por esto que para saber en qué medida un rasgo latente está presente en las personas, los investigadores e investigadoras definen las operaciones mediante las cuales se puede determinar su presencia o ausencia (o magnitud). Esto se conoce como operacionalización de variables.

Por lo general, estas operaciones son mediciones que involucran registros numéricos (como los minutos que invierten las personas en resolver una tarea o sus puntuaciones en una escala de motivación de logro). De hecho, las escalas o test psicométricos se definen comúnmente como un conjunto de indicadores operacionales de un rasgo o dominio de comportamiento (Tornimbeni, Pérez & Olaz, 2008).

En la medida en que operacionalizar implica definir la presencia, ausencia o magnitud de los fenómenos en términos numéricos, es preciso introducir algunas

nociones básicas sobre medición y psicometría, especialmente los conceptos de validez y confiabilidad.

Los conocimientos actuales sobre estos tópicos provienen de tres grandes enfoques o teorías: la Teoría Clásica de los Tests (TC), la Teoría de la Generalizabilidad (TG) y la Teoría de la Respuesta a los Ítems (TRI). Aquí no vamos a tratar estas teorías, de modo que ofrecemos al final del capítulo recomendaciones sobre textos para profundizar en ellas. Nuestro interés principal es describir sucintamente los conceptos de validez y confiabilidad, dar a conocer las principales estrategias utilizadas en la investigación moderna para estimarlas y de esta manera comprender el importante papel que juega la EC en todo este complejo proceso.

Medición y Psicometría

Medir, como ya lo indicamos, es un procedimiento de asignación de números a propiedades, de modo tal que los números las caractericen (Martínez, 1996). El tema de la medición implica entonces derivar reglas de asignación numéricas; esto es, generar las correspondencias entre un sistema empírico y un sistema numérico. El área de la Psicología que se encarga de todos estos asuntos es la Psicometría.

Para Galton (1879) la psicometría consiste en datar el origen de los pensamientos de una manera matemática y fiable. A su vez Martínez (1996) establece que la Psicometría es la “construcción de instrumentos que sirvieran para la asignación de números a atributos o conductas de las personas” (pp 21-22). Y siguiendo esa lógica, McIntire y Miller (2007), así como Kaplan y Sacuzzo

(2006), definen la Psicometría como aquellos aspectos técnicos y cuantitativos de la medición de eventos cognoscitivos, conductuales y emocionales.

Independientemente de las formas de recolección de información, pero particularmente cuando se trata de instrumentos de autoreporte, la medición trae consigo dos preguntas fundamentales: a) ¿en qué medida los instrumentos miden lo que pretenden medir? y b) ¿en qué medida son tales mediciones de fiar?

En efecto, cualquier medición captura mucho más de lo que se pretende medir, como otros constructos que no son de interés y por supuesto errores de medición aleatorios. La Psicometría trata precisamente de proporcionar las herramientas para determinar en qué medida nuestros instrumentos miden consistentemente (confiabilidad) lo que tienen que medir (validez).

Validez

Siguiendo a Nunnally (1991) podemos decir que las puntuaciones de un instrumento poseen propiedades de validez cuando es posible obtener evidencia de que el test mide lo que pretende medir y no otra cosa, justificando así las inferencias que podemos hacer sobre sus resultados.

En los enfoques modernos no se conceptualiza la validez como una característica inherente al instrumento, sino una propiedad del significado que podemos darle a las puntuaciones obtenidas mediante este y las consecuencias de las interpretaciones de tales puntuaciones derivadas de él (Messick, 1989, 1995). La interpretación o significado de las puntuaciones depende no solo de los reactivos, sino también de las personas a las que aplicamos el instrumento, el

contexto de aplicación y el modelo teórico que define el constructo en estudio y sus relaciones con otros constructos.

Por lo tanto, estimar la validez de las interpretaciones es un problema que requiere de un esfuerzo científico igual al que se exige para examinar cualquier hipótesis de investigación, en otras palabras, en el campo de la medición psicológica es imprescindible proporcionar evidencias empíricas que apoyen o refuten cualquier inferencia (Cronbach & Meehl, 1955).

La validez no existe en términos absolutos, es decir, no podemos decir que un instrumento sea válido o inválido. La validez de las inferencias aumenta o disminuye, es relativamente robusta o frágil, dependiendo de las evidencias empíricas. Dentro de los estándares modernos de medición psicológica y educativa (AERA, APA & NCME, 1999) se incluyen varios tipos de evidencias empíricas que podemos recopilar para estimar la validez de las inferencias que hacemos a partir de nuestros instrumentos. Estas evidencias se basan en: a) el contenido de los reactivos, b) el proceso de respuestas al instrumento, c) la estructura interna del test, d) las asociaciones de las puntuaciones con los puntajes de variables externas al instrumento y e) las consecuencias de su aplicación.

Sobrepasa los objetivos del presente texto hacer una exposición extensa sobre cada uno de los aspectos involucrados en el proceso de producir evidencias de cada una de estas dimensiones. Solo vamos a ofrecer una descripción muy básica de ellas. En el Cuadro 1 presentamos los cinco tipos de evidencia de validez, las interrogantes que pretende responder y algunos de los

procedimientos empíricos que normalmente se recomiendan para recopilar dichas evidencias. Al final del capítulo, recomendamos algunos textos y autores clave para profundizar en estos temas.

Cuadro 1. Dimensiones de Validez

Dimensión	Interrogante clave	Estrategias o técnicas analíticas
CONTENIDO	¿Existe correspondencia entre el contenido del instrumento y el dominio que pretende medir?	Acuerdo entre jueves, Kappa de Cohen, correlación intra-clase
PROCESO PREGUNTA/ RESPUESTA	¿Concuerda la naturaleza y el proceso de la respuesta del individuo con el constructo que se pretende medir?	EC, panel de expertos
ESTRUCTURA	¿Corresponde la covariación de los reactivos a las dimensiones que se presenten medir?	Análisis de factores tanto exploratorios como confirmatorios
CONVERGENTE/ DISCRIMINANTE	¿Convergen las puntuaciones con otras medidas del mismo constructo?, ¿se diferencian los puntajes con respecto a medidas que miden constructos distintos?	Correlaciones bivariadas, multirasgo-multimétodo
EN RELACIÓN CON CRITERIOS EXTERNOS	¿Permiten las puntuaciones del test predecir criterios externos a la ejecución del propio instrumento?	Pruebas de hipótesis sobre dos promedios, análisis de varianza, correlación bivariada, regresión múltiple, modelos de ecuaciones estructurales, etc.
CONSECUENCIAS DE LA APLICACIÓN	¿Las decisiones que se toman a partir de los resultados del test tienen efectos adversos para los sujetos o grupos de sujetos?	Análisis del funcionamiento diferencial del ítem (DIF)

Las evidencias basadas en el contenido del test ofrecen información para determinar en qué medida los reactivos del test representan, abarcan, cubren o son representativos del dominio que pretenden medir. Una manera común de

recopilar evidencia sobre la validez del contenido consiste en presentar los reactivos a jueces expertos para que determinen la concordancia esperada entre el contenido del test y el constructo. Por lo general se mide el grado de acuerdo entre jueces y se analizan las fuentes de discordancia (AERA, APA & NCME, 1999).

La evidencia basada en el proceso de respuesta se obtiene examinando los mecanismos mentales o cognitivos que los sujetos ejecutan para responder a los reactivos, con el objetivo de determinar si los sujetos están realizando las operaciones necesarias para que los investigadores puedan inferir la presencia, ausencia o nivel del constructo que están midiendo. Por lo general, esta evidencia se obtiene consultando directamente a los sujetos, mediante entrevistas cualitativas, protocolos de pruebas cognitivas o cualquier otro procedimiento para evaluar las estrategias de respuesta a los reactivos (Cortada de Kohen, 2005; Embretson & Gorin, 2001).

Las evidencias basadas en la estructura son aquellas que obtenemos al mostrar que la covariación empírica de los reactivos refleja la estructura que se supone debe tener el instrumento; por ejemplo, si este fue diseñado para medir una sola dimensión, es de esperarse que los reactivos que miden ese constructo formen un solo factor o dimensión. Por lo general se utiliza el análisis de factores (exploratorio y confirmatorio) para determinar la estructura subyacente a la covariación empírica de los reactivos como evidencia de validez estructural (Muñiz, 2002).

Las evidencias basadas en las asociaciones del instrumento con variables externas pueden especificarse en dos tipos, validez convergente-discriminante y validez de criterio. Para estimar la validez convergente de una medición, el procedimiento razonable consiste en incluir otras mediciones que miden el mismo constructo y determinar hasta qué punto estas covarían, pues si miden lo mismo deberían correlacionarse positivamente. Para determinar la validez discriminante, el procedimiento común es incluir mediciones que miden constructos distintos al constructo de interés y determinar cuánto covarían, bajo el supuesto de que si no miden el constructo de interés entonces las mediciones deberían estar poco correlacionadas (Tornimbeni et al., 2008).

Para recopilar evidencias sobre la validez de criterio se utilizan procedimientos similares a los anteriormente descritos, solo que aquí se emplea un criterio externo a la ejecución del test, como pueden ser las medidas de alguna variable que el instrumento intenta predecir (Cronbach & Mehl, 1955). Así, en la evaluación de una medida de *burnout* o estrés laboral, un investigador puede recopilar información sobre la capacidad del test para predecir la cantidad de ausencias injustificadas o la cantidad de solicitudes de incapacidad laboral, bajo el supuesto (derivado de su teoría) de que altos niveles de *burnout* predicen altos niveles de ausentismo. Variables categóricas externas, como el sexo o la pertenencia a grupos específicos (esquizofrénicos vs. no esquizofrénicos), también permiten a los investigadores estimar la validez de las medidas si al comparar las puntuaciones del test los grupos difieren significativamente entre ellos y de la manera en que la teoría plantea que deberían hacerlo.

Para todos estos casos, la estadística moderna ofrece un importante número de estrategias analíticas, como la estimación de las correlaciones bivariadas entre las variables en estudio, la comparación de promedios (pruebas *t* y análisis de varianza) y análisis multivariados más sofisticados, como el análisis de varianza múltiple, el análisis de regresión múltiple, los modelos de ecuaciones estructurales y los modelos lineales jerárquicos.

La validez de las consecuencias hace referencia a la necesidad de examinar los potenciales efectos colaterales no anticipados de los usos legítimos del test, derivados de fuentes de invalidez del instrumento (Messick, 1995). Este tipo de validez es particularmente importante para las pruebas de ejecución máxima (rendimiento, aptitudes, inteligencia) y las "de alto riesgo" (selección, promoción o certificación), es decir, aquellas en las que los resultados tienen consecuencias importantes para los sujetos o grupos de sujetos, como por ejemplo las pruebas cuyos resultados definen el ingreso de los aplicantes a centros educativos, o aquellas empleadas para la selección de personal y la evaluación de la habilidad mental de las personas para ejecutar alguna tarea (portar armas, por ejemplo) (Padilla, Gómez, Hidalgo & Muñíz, 2006). Así, si existen evidencias de que un instrumento favorece a un grupo de aplicantes sobre otro, el test pierde una importante parte de su validez, pues no cumple con su finalidad en el marco de los principios de justicia en la medición.

Existen varios procedimientos para analizar las consecuencias de los resultados de un instrumento. Dentro de ellos destaca el Análisis Diferencial del Ítem (DIF, por sus siglas en inglés), el cual es un procedimiento muy utilizado para

detectar si un reactivo favorece a un grupo de aplicantes sobre otro (Moreira, 2008). En una prueba de razonamiento matemático, por ejemplo, un reactivo presenta DIF cuando examinados que poseen el mismo nivel en el rasgo medido (razonamiento matemático) presentan diferentes probabilidades de acertar el reactivo solo por pertenecer a distintos grupos (ser mujer, por ejemplo) y no porque difieren en el nivel de constructo medido. Las diferentes probabilidades se calculan comparando las frecuencias de aciertos y errores en un ítem en los sujetos que, perteneciendo a distintas poblaciones, muestran el mismo nivel de puntuación en la prueba (Morerira, 2008).

Confiabilidad

Anastasi (1982) define la confiabilidad de un instrumento como la precisión con que el test mide lo que mide, en una población determinada y en las condiciones normales de aplicación.

Otra forma de entender la confiabilidad es en términos de la consistencia entre las puntuaciones que obtuvieron las mismas personas en momentos distintos: si un test es aplicado a las mismas personas (pongamos por caso, dos) en ocasiones distintas, se esperaría que las puntuaciones de las dos aplicaciones arrojen resultados similares, ya que están midiendo lo mismo en las mismas personas. También puede verse como la consistencia entre las puntuaciones que obtienen los mismos sujetos en dos conjuntos de ítems equivalentes, porque si los dos conjuntos de estos son equivalentes, se esperaría que las puntuaciones de ambos fuesen similares (AERA, APA y CMRE, 1999).

El hecho de que la confiabilidad apele a la precisión nos recuerda que la confiabilidad indica hasta qué punto la medida no contiene errores de medición. El axioma fundamental de la Teoría Clásica de los Tests ($X_i = T_i + E_i$) presupone que en toda medición están involucrados un valor verdadero del constructo y errores de medición aleatorios. La T (*true*) representa la puntuación media de un número infinito de mediciones en un sujeto y la E (*error*) representa todas las variaciones puntuales en la medición que no están relacionadas con el procedimiento de medición. Tales errores se suponen aleatorios, lo que permite la definición del error en términos estadísticos (Muñiz, 2002; Osterlind, 2006).

Así, si una medición contiene pocos errores aleatorios, entonces las puntuaciones de esta pueden ser reproducidas si el constructo es medido de nuevo utilizando el mismo instrumento en los mismos sujetos: si los errores son verdaderamente aleatorios, tendrían muy baja probabilidad de aparecer nuevamente.

Nótese que hemos hecho énfasis en definir los errores como aleatorios. En realidad existen dos tipos de errores: los errores constantes o sistemáticos (muchas veces conocidos como sesgos) y los errores aleatorios. Los primeros se producen cuando las puntuaciones obtenidas con una escala son sistemáticamente mayores o menores que lo que realmente deben ser cuando se aplica la escala. Los segundos (aleatorios o no sistemáticos) se observan cuando las puntuaciones son algunas veces mayores y otras veces menores de lo que realmente deben ser, pero sin un patrón sistemático.

Al igual que en el caso de la validez, la confiabilidad no es absoluta y depende no solo del contenido de los reactivos, sino también de las personas que toman el test, las condiciones de aplicación del instrumento, las personas que lo califican y las formas de calificarlo. Asimismo, aquí el oficio del constructor o constructora de instrumentos no es distinto del científico o la científica que debe proporcionar evidencias empíricas a favor del argumento de que sus instrumentos son confiables, tomando en cuenta todos los aspectos que intervienen en la aplicación que representan potenciales fuentes de error.

En general, se distinguen tres grandes dimensiones de confiabilidad: a) la estabilidad de las puntuaciones, b) la consistencia interna y c) la congruencia entre calificadores (algunas veces referida como “objetividad”) (Tornimbeni et al., 2008).

Para estimar la estabilidad de una medición, el procedimiento elemental consiste en medir el constructo con el mismo instrumento en las mismas personas en dos momentos distintos y calcular la correlación de las puntuaciones observadas en ambas ocasiones. Un coeficiente de correlación alto entre los dos grupos de puntuaciones indicaría que los individuos mantuvieron sus posiciones en las dos aplicaciones. Esta estrategia se conoce como método test-retest.

También se puede estimar la estabilidad de un test administrando dos (o más) formas paralelas de la misma medición en dos ocasiones distintas. Las formas paralelas deben tener las mismas características formales y estadísticas; entre otros requisitos, deben contener la misma cantidad de reactivos y la misma escala de medida, además de tener medias y desviaciones estándar semejantes. Esta estrategia es conocida como el método de formas paralelas.

Cuando las dos versiones equivalentes son aplicadas en una sola ocasión, la correlación entre los dos conjuntos de ítems es un indicador de consistencia interna. Otra forma de estimar la consistencia interna es el método de mitades equivalentes (split-half). Una vez aplicado el test, éste se divide en dos mitades (por ejemplo, los reactivos pares vs. los reactivos impares) y se califican por separado; finalmente, se calcula la correlación entre las dos series de puntajes resultantes.

Esta idea de partir el test por la mitad es la base de los llamados métodos de covarianza de los ítems o métodos de equivalencia racional (Anastassi, 1982). En dichos métodos se considera que si un test está formado por un conjunto de *x* ítems, estos pueden ser considerados como un conjunto de *x* test paralelos (tantos como ítems tenga el instrumento). Luego se deriva una ecuación para calcular el coeficiente de consistencia interna, que básicamente es el promedio de correlaciones de todos los test paralelos (o reactivos). Los coeficientes más utilizados son el (famoso) coeficiente Alfa de Cronbach y el coeficiente Kuder-Richardson (para reactivos dicotómicos, como los de “falso y verdadero”) (Osterlind, 2006).

Para estimar el acuerdo entre calificadores se aplica la prueba una sola vez y se entregan los resultados a los jueces, quienes los califican de manera independiente, para luego estimar el grado de acuerdo en la calificación. Este procedimiento es particularmente idóneo para instrumentos que se deben calificar de manera objetiva. El objetivo de este procedimiento es que el test produzca los mismos resultados, independientemente de la subjetividad del evaluador, de lo contrario no serían precisos. Los índices de acuerdo más utilizados son los

coeficientes Kappa (para variables nominales), las w de Kendall y la correlación intraclase (para variables medidas en el nivel ordinal o de intervalo) (Tornimbeni, et al., 2008).

Un análisis complementario al de la consistencia interna del test o instrumento como totalidad se conoce como análisis de ítems, mediante el cual es posible analizar las características de los ítems directamente vinculadas con las propiedades del test y que por tanto influyen en su precisión (Muñiz, 2002). Las principales características son la dificultad del reactivo y su capacidad de discriminación.

El índice de dificultad se calcula en aquellos reactivos en donde existe una respuesta correcta, y se define como la proporción de individuos que aciertan el reactivo. El índice de discriminación informa sobre la capacidad del reactivo para distinguir (diferenciar) entre quienes tienen puntuaciones altas en la escala o test y quienes tienen puntuaciones bajas, y se define como la correlación simple entre las puntuaciones de los sujetos en el reactivo y sus puntajes en la escala como un todo (Muñiz, 2002).

En el Cuadro 2 se resumen las diferentes dimensiones de la confiabilidad, así como diversas técnicas a las que se puede acudir para su verificación.

Como se puede observar a lo largo de esta breve exposición, estimar la validez y la confiabilidad forma parte de los requerimientos básicos de desarrollo de instrumentos. Se trata de un proceso idéntico al que se realiza en cualquier investigación empírico-analítica: El investigador o investigadora formula un argumento o hipótesis sobre las distintas facetas o dimensiones de validez y

confiabilidad, y ejecuta uno o varios estudios para recopilar evidencias que apoyen (o refuten) su argumento. Dentro de este proceso, las fuentes de información son tanto cualitativas como cuantitativas.

Cuadro 2. Dimensiones de Confiabilidad

Dimensión	Método	# de aplicaciones	Estadísticos Más usados
Estabilidad	Test-retest	2	*r*
	Formas paralelas	2	*r*
Consistencia interna	Formas paralelas	1	*r*
	Partición por mitades	1	*r* (con corrección Spearman-Brown
	Covariación entre reactivos	1	α de Cronbach, Kuder-Richardson
Objetividad	Acuerdo entre examinadores	1	Kappa de Cohen, *w* de Kendall, coeficiente de correlación intraclase

Nota. r = Coeficiente de correlación de Pearson.

Es particularmente obvio el papel de la EC dentro de las estrategias para recolectar evidencias de validez como herramienta central para el análisis del proceso de respuesta. Este tipo de análisis tiene como objetivo permitir a los(as) investigadores(as) examinar en qué medida su instrumento mide lo que debe medir, analizando la naturaleza del proceso mental de responder los reactivos. Figura dentro de los primeros pasos en la construcción, adaptación o validación de

los instrumentos de medición y normalmente este tipo de evidencia se recopila antes de probar el instrumento en el estudio piloto, donde es necesaria una muestra grande para recabar los otros tipos de evidencias de validez.

La EC refiere precisamente a la evaluación de los mecanismos cognitivos que se encuentran a la base de responder los reactivos en ambientes controlados con una modesta muestra de potenciales aplicantes, entrevistados o examinados, y por ello es una de las estrategias privilegiadas por los investigadores para recopilar evidencias de este tipo de validez.

Aunque la EC no es la única estrategia para recopilar información sobre el proceso de repuesta, es sin duda alguna el dispositivo idóneo para hacerlo, precisamente porque su objetivo es estudiar en el laboratorio el funcionamiento de los reactivos desde la perspectiva de los participantes antes de probarlos en muestras más grandes.

La EC, como estrategia para estudiar el proceso de respuesta, posibilita no sólo tener certeza sobre lo que significan las puntuaciones de la medida, sino también conocer más a fondo el constructo que se está midiendo y detectar las potenciales fuentes de invalidez (como por ejemplo, el funcionamiento diferencial de los reactivos), esclareciendo en qué medida el reactivo demanda la puesta en marcha de otras capacidades irrelevantes o evoca tendencias no deseadas, como la de conformarse con las normas.

En el caso del análisis de confiabilidad, la EC no aparece claramente perfilada, pero parece razonable pensar que la EC tiene una importante utilidad para la precisión de los instrumentos. Consultar directamente a los potenciales

evaluados sobre los procesos mentales que ponen en marcha a la hora de contestar los reactivos proporciona información sobre la dificultad de los reactivos.

La incorporación de la evidencia basada en el proceso de respuesta es producto de los aportes de las Ciencias Cognitivas a la Psicometría. En la siguiente sección describimos algunos de esos aportes al concentrarnos en los principios teóricos de la EC.

Literatura recomendada

Son especialmente claves los textos clásicos de Anastassi (1982), Cronbach y Meehl (1955), Messick (1989, 1995) y Nunnaly (1991). De los autores modernos recomendamos los trabajos de Susan Embretson (Embretson & Gorin, 2001), porque incorporan sistemáticamente los aportes de las Ciencias Cognitivas a la Psicometría y los trabajos de Bollen (2002) que son particularmente iluminadores sobre el concepto de variables latentes.

En el contexto costarricense existen varios trabajos que pueden ser de gran utilidad para profundizar en todos estos temas. Por ejemplo, en el texto de Eiliana Montero del 2008 se discute la diferencia entre índices y escalas y se ofrecen ejemplos concretos sobre su construcción (Montero, 2008). En el mismo año, Tania Moreira publicó un artículo en donde describe muy claramente el concepto de DIF y sus aplicaciones (Moreira, 2008).

No tratamos aquí las teorías actuales que guían la Psicometría. En su lugar recomendamos el texto de Muñiz (2002) sobre Teoría Clásica de los Tests, pero pensamos que es especialmente idóneo para los lectores y lectoras más

experimentados. Para los menos experimentados recomendamos el texto de Tornimbeni, Pérez y Olaz (2008), quienes hacen una introducción a la Psicometría muy clara y accesible. Por su parte María Elena Zúñiga y Eiliana Montero (2007) hacen una excelente exposición de la Teoría de la Generalizabilidad, e Eiliana Montero (2001) hace lo propio con la Teoría de Respuesta Ítem.

Todos estos textos figuran en la lista de referencias al final de este cuaderno y muchos de ellos están disponibles en la Web.

LAS CIENCIAS COGNITIVAS Y EL DISEÑO DE INSTRUMENTOS DE PAPEL Y LÁPIZ

La preocupación por la evaluación del funcionamiento de los reactivos no es nueva en las ciencias que utilizan instrumentos de papel y lápiz como medio para recolectar información. Desde muy temprano, los y las investigadores(as) se preocuparon por que la redacción de los estímulos permitiera a los sujetos dar cuenta de sus experiencias, actitudes y sentimientos. Así por ejemplo, en el marco del estudio de las actitudes, Thurstone (1928) recomendaba:

> En la elaboración de la lista inicial de aseveraciones se aplican varios criterios prácticos en el primer trabajo de edición. Algunos de los criterios más importantes son los siguientes: A) las afirmaciones deben ser lo más cortas posible de manera que no fatiguen a los sujetos a los que se les pide leer la lista completa. B) Las afirmaciones deben ser de tal tipo que puedan ser secundadas o rechazadas conforme a su concordancia o discrepancia con la actitud del lector […] C) Cada aseveración debe prever que su aceptación o rechazo indique algo con respecto a la actitud del lector acerca del asunto en cuestión […] D) Las afirmaciones de doble significado deben evitarse […] Las afirmaciones de doble significado tienden a tener un alto componente de ambigüedad […] E) Es necesario asegurarse de que por lo menos la gran mayoría de las afirmaciones pertenecen realmente a la variable de actitud que se va a medir (pp. 544-545, traducción nuestra).

Estas y otras reglas, de conocimiento generalizado entre quienes construyen instrumentos de papel y lápiz, coinciden con los principios básicos de

la comunicación y cognición humana. Pero no fue sino hasta los años 70, coincidiendo con el giro cognitivo en la Psicología, que se empezaron a realizar esfuerzos sistemáticos por incorporar los hallazgos de las ciencias cognitivas en la construcción de reactivos.

Uno de los esfuerzos más notables de esta cooperación interdisciplinaria se dio a partir de 1978 entre académicos de las Ciencias Cognitivas e investigadores de las grandes casas encuestadoras de los gobiernos de países como Estados Unidos, Gran Bretaña y Alemania. De esta cooperación emergen los llamados métodos cognitivos para la evaluación de las encuestas y los formularios de encuesta, de los cuales proviene la EC (Jobe & Mingay, 1991).

Mucha de la investigación en esta línea se preocupó por estimar el efecto del diseño de los instrumentos, como la redacción, el orden de presentación de los reactivos y el formato de las escalas de respuesta, en las reacciones de los y las entrevistados (Schwartz & Oyserman, 2001).

Otra área de investigación se focalizó en la memoria autobiográfica, específicamente en los procesos cognitivos que se activan cuando los encuestados responden preguntas sobre sí mismos y sus experiencias pasadas (Tourangeau, Rips & Rasinski, 2004).

Otros estudios se han ocupado de los procesos de estimación de la respuesta, es decir, los heurísticos utilizados por los encuestados para estimar una respuesta adecuada a las preguntas que se le realizan (Job & Mingay, 1991).

Finalmente, existe todo un trabajo alrededor de tópicos sensibles, como el reporte de conductas socialmente reprochables y el impacto del contexto de

aplicación de las encuestas, el formato de respuesta y el tipo de método de recolección de datos en la respuesta a este tipo de preguntas (Tourangeau & Smith, 1996).

Los resultados de investigación de estas líneas de trabajo se han utilizado para derivar distintos modelos teóricos sobre los mecanismos cognitivos implicados en el proceso de responder instrumentos de papel y lápiz. Estos conocimientos han resultado de gran utilidad para evaluar el funcionamiento de los reactivos y mejorarlos, y son las bases teóricas que guían la EC. En la presente sección nos concentraremos en los principios teóricos y las evidencias empíricas que han aportado las Ciencias Cognitivas para comprender mejor los procesos mentales involucrados a la hora de contestar instrumentos de medición.

Las etapas del proceso pregunta-respuesta

La EC se basa en la idea de que para contestar un cuestionario las personas primero deben comprender el significado transmitido en este. Posteriormente, inician el proceso de recuperar la información solicitada de la memoria autobiográfica, luego recurren a diversas estrategias para estimar la respuesta adecuada y finalmente emiten la respuesta elegida.

Este modelo general se conoce como el Modelo Tourangeau de cuatro etapas, en alusión a su autor (Tourangeau, 1984), y las cuatro etapas se denominan a) comprensión, b) recuperación, c) estimación y d) ejecución de la respuesta. Se trata de un modelo muy simple, pero su parsimonia resulta de gran utilidad para detectar y corregir potenciales problemas en los instrumentos de papel y lápiz. En la Figura 1 se presenta una descripción esquemática del modelo.

Como se puede observar, el modelo no supone que las etapas se suceden de manera lineal, sino que más bien las concibe como procesos interrelacionados a los cuales el sujeto recurre en distintos momentos dependiendo de la finalidad de su acción. Por ejemplo, un sujeto puede volver a leer una pregunta y redefinir su significado en el momento en que mira las opciones de respuesta y puede recuperar información distinta en función de la forma en que pretende contestar el reactivo.

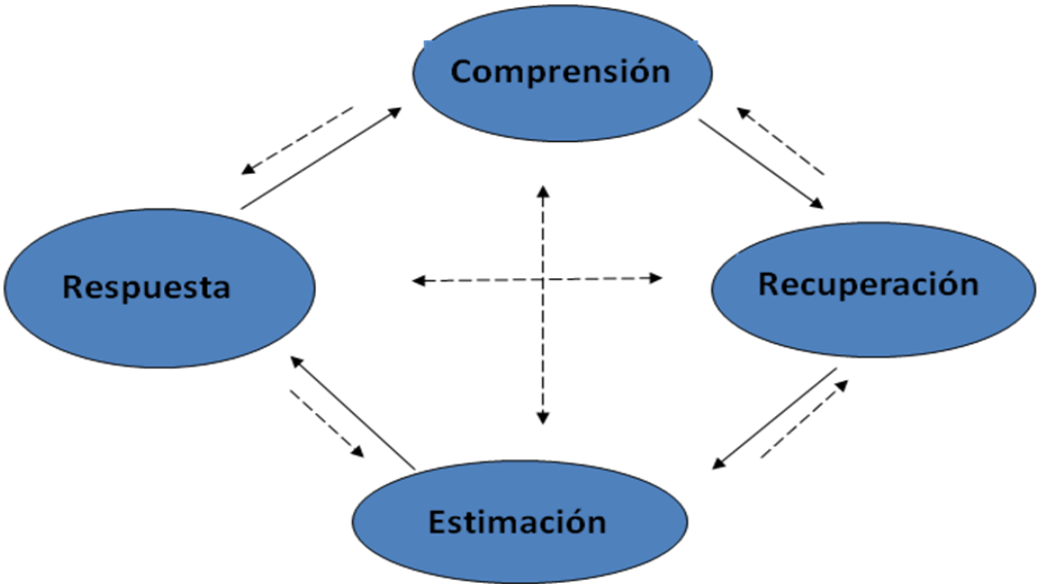


Figura 1. Modelo de cuatro etapas.
(Adaptado de Tourangeau, 1984)

Ahora bien, por más sencillo que este proceso parezca, la investigación en comunicación, lenguaje, memoria, aprendizaje y motivación nos recuerda que una pregunta tan simple como “En los últimos 12 meses, ¿cuántas veces fue usted al

dentista?" puede resultar bastante ambigua y puede ser contestada de muchas maneras dependiendo de múltiples factores.

Más aún, pese a que reportar conductas y actitudes relevantes toma un considerable tiempo, las encuestas no suelen usar entrevistas a profundidad, sino estandarizadas, que por lo general son aplicadas en pocos minutos y en muchos casos se llevan cabo mediante cuestionarios autoaplicados. Para complicar más la situación, el contexto de aplicación y el cuestionario en sí mismo proporcionan información a los encuestados y encuestadas muchas veces irrelevante para los fines de la investigación, lo cual puede desviar su atención a otros aspectos, limitar sus estrategias para recuperar la información solicitada y, en el peor de los escenarios, influir en sus respuestas de manera no deseada.

De allí la necesidad de conocer más a fondo lo que acontece mentalmente en cada etapa del proceso de responder un reactivo, con el fin de incorporar este conocimiento en la evaluación de los mismos. A continuación describimos de manera muy resumida algunos elementos de los procesos mentales involucrados en cada una de las etapas propuestas por Tourangeau (1984).

Comprensión

La investigación nos recuerda que la comprensión del reactivo no refiere únicamente a la capacidad de los sujetos para entender la pregunta que se les realiza, refiere también a las distintas formas de interpretar un reactivo (Schwarz & Oyserman, 2001). Comprender las palabras contenidas en los reactivos, o su sentido literal, no significa que las personas comprendan las intenciones

comunicativas (o sentido pragmático) de los y las investigadores (Tourangeau, et al. 2004).

La investigación indica que en las encuestas los y las participantes infieren el significado e intención de las preguntas utilizando las normas tácitas que normalmente utilizan en las conversaciones cotidianas. Una de esas normas es el Principio de Cooperación, según la cual la comunicación se basa en las siguientes máximas: a) proporcionar la información relevante y vinculada a lo que se solicita, b) no brindar ni más ni menos información de la que se solicita, c) transmitir contenidos claros, evitando ambigüedades, d) evitar expresar contenidos que se consideren falsos y e) evitar sostener afirmaciones de las que no se tengan pruebas adecuadas (Schwarz & Oyserman, 2001).

En la vida real, las personas tienen claves contextuales para responder a las demandas de información (el lenguaje no verbal, las experiencias previas de interacción con el interlocutor, los estereotipos sobre los otros, etc.) y de esa manera saben intuitivamente cómo aplicar estas máximas de comunicación. En el mundo de los cuestionarios de papel y lápiz, el contexto de donde las personas infieren los requerimientos de información es el cuestionario mismo y la situación de entrevista. Los entrevistados monitorean e infieren del cuestionario o la entrevista lo que se desea de ellos para encontrar las claves contextuales de comunicación (Willis, 2005).

En un cuestionario, las claves contextuales son de diversa naturaleza, como el formato de respuesta, la ubicación de las preguntas en el cuestionario, los períodos de referencia (lugar y tiempo) de la información solicitada, la forma de

aplicación de los instrumentos y hasta la afiliación de los investigadores (Schawarz & Oyserman, 2001). Veamos un par de ejemplos.

En un estudio realizado en la Universidad de Michigan, Norenzayan y Schwarz (1999) solicitaron a 60 participantes leer un caso de asesinato. A la mitad de los participantes se les indicó que el estudio era llevado a cabo por el "Instituto de Investigación en Personalidad" y a la otra mitad se le indicó que el estudio era llevado a cabo por el "Instituto de Investigación Social". La clave sobre la afiliación de los investigadores aparecía en los cuestionarios en donde aparecía la historia; después de leerla, a los participantes se les solicitó dar al menos cinco razones por las cuales ellos creían que la persona cometió el asesinato.

Cuando el cuestionario estaba a nombre del Instituto de Investigación en Personalidad, las explicaciones de los participantes se focalizaban en características de personalidad del perpetrador (atribuciones internas y estables de los asesinos), mientras que cuando el mismo cuestionario estaba a nombre del Instituto de Investigación Social, las explicaciones se concentraron en los determinantes sociales del homicidio (atribuciones externas). Esto se observó independientemente del lugar en donde estaba localizada la información sobre el instituto (antes o después de leer la historia) y del tipo de delito (tiroteo o atentado con bomba).

Los resultados sugieren que los encuestados adoptaron el principio de proporcionar la información más relevante, adecuada y vinculada al contexto y las características del interlocutor: A los "investigadores" del Instituto de Investigación Social les ofrecieron más explicaciones sociales sobre los crímenes y a los "investigadores" del Instituto de Investigación en Personalidad, más explicaciones

sobre las características de personalidad de los perpetradores. No en vano el artículo que describe esta investigación fue titulado por sus investigadores de la siguiente manera: “Diciendo lo que ellos quieren saber: los participantes ajustan sus atribuciones causales a los intereses de sus investigadores” (Norenzayan y Schwarz, 1999, p. 1011, traducción nuestra).

Otro ejemplo sobre el uso de claves contextuales proviene de un estudio del Centro Nacional de Estadísticas en Salud de los Estados Unidos realizado por Lessler, Tourangeau y Salter (1989). Ellos compararon el reporte de las visitas al dentista utilizando dos distintos tipos de formato. El formato estándar preguntaba directamente a las personas sobre la frecuencia con que los miembros de su familia fueron al dentista en los últimos 12 meses. El formato experimental preguntó a los participantes lo mismo, pero antes le ofreció una lista de posibles razones por las que las personas van al dentista (ver Cuadro 3).

Cuadro 3. Dos versiones de una misma pregunta

Versión Experimental

Las siguientes preguntas tienen como objetivo saber cuántas veces cada miembro de su familia ha ido al dentista en el último año. Para ayudarle a recordar las posibles visitas le voy a leer una lista de razones por las cuales las personas pueden ir al dentista:

1. Algunas personas van al dentista para un chequeo, limpiarse los dientes o ponerse una calza
2. Algunas personas van por que les duelen los dientes o porque se les cayó una calza
3. Algunas personas van como parte de un tratamiento particular como tratamiento del nervio
4. Y algunas personas van como parte de un tratamiento de ortodoncia (frenillos)

En los últimos 12 meses (desde hace un año), ¿cuántas veces _____ fue al dentista?

Versión Estándar

En los últimos 12 meses (desde hace un año), ¿cuántas veces _____ fue al dentista?

Nota. Traducido y adaptado de Lessler, Tourangeau y Salter (1989).

Los resultados de la comparación de ambos formatos indicaron que con la versión ampliada (con ejemplos) las personas tendían a reportar más visitas al dentista que con el formato estándar (2.1 visitas al año con la versión ampliada vs. 1.2 visitas al año con el formato estándar). Durante el análisis del proceso pregunta respuesta mediante EC, los investigadores notaron que ante la pregunta estándar muchos de los entrevistados entendían que debían responder cuántas veces al año "se supone" que uno debe ir al dentista (aproximadamente dos).

Al ofrecerles las diversas razones por las cuales una persona puede ir al médico, los investigadores hicieron explícita la intención de su pregunta, como diciendo "nos interesan todas las visitas al dentista que hizo en los últimos 12 meses, no importa la razón, la costumbre o la norma". Un reactivo así formulado no sólo permitió a las personas tener claridad sobre la intención de la pregunta, sino que también ayudó a recuperar la información deseada, que es quizás el proceso más complejo de todas las etapas del proceso pregunta-respuesta, como veremos a continuación.

Recuperación de la información

Una vez comprendida la pregunta, los encuestados deben recuperar la información de la memoria. Como ya indicamos, este es quizá uno de los temas más controversiales a la hora de plantear preguntas en instrumentos de papel y lápiz. De allí que la mayoría de la investigación se haya concentrado en este tema.

Las teorías modernas distinguen entre dos tipos de conocimiento almacenado en la memoria: el declarativo y el procedural (Klein, German, Cosmides & Gabriel, 2004). El conocimiento declarativo es el que se refiere a la

información sobre hechos, mientras que el procedural, al repertorio de reglas y habilidades que nos permiten navegar en el mundo. El primero se relaciona con el "saber qué" (saber sobre las cosas en el mundo y sobre el sí mismo) y el segundo, al "saber cómo" (saber sobre cómo actuar frente al mundo y los otros) (Ryle, 1949).

La memoria declarativa, por su parte, parece tener dos formas básicas: episódica y semántica (Tulving, 2002). La episódica consiste en el conocimiento de los eventos que nos han sucedido en nuestras experiencias pasadas, junto con la conciencia de que sucedieron. Se trata de una experiencia fenoménica personal, el sentimiento de que eso sucedió y nos sucedió a nosotros: es un conocimiento que uno puede "revivir". La memoria semántica, por su parte, no necesariamente está acompañada de la conciencia de que "eso me sucedió a mí". Refiere más bien al conocimiento sobre conceptos y significados (como "perro", "calor", "uno", "dos", "tres") que han sido obtenidos sin referencia a dónde y cuándo lo hemos adquirido. En las encuestas solicitamos información de ambos tipos, pero la memoria episódica es particularmente solicitada.

De acuerdo con autores como Tulving (2002) o Klein, German, Cosmides & abriel (2004), para que la memoria sea experimentada como conocimiento autobiográfico, se requieren al menos tres capacidades: a) la autorreflexión, es decir, la habilidad de reflexionar sobre nuestros estados mentales: saber sobre lo que sabemos; b) la noción de pertenencia de nuestros recuerdos: la sensación de que nosotros somos la causa de nuestros pensamientos y acciones, junto con el sentimiento de que los pensamientos y acciones nos pertenecen; y c) el

reconocimiento de la temporalidad personal, esto es, poder “navegar” en el propio tiempo como una sucesión de experiencias personales.

La investigación sobre estos tópicos nos indica que todavía nos falta mucho por conocer sobre la memoria y que en algunos puntos no existe un consenso generalizado entre los investigadores. No obstante, existe mucha evidencia a favor de cuatro características básicas de estos tipos de memoria, con implicaciones muy importante para quienes diseñan cuestionarios.

La primera característica refiere al hecho de que la memoria autobiográfica decrece con el tiempo y que la recuperación depende de las características temporales de los eventos. La segunda es la observación de que la memoria episódica no siempre coincide con la memoria semántica. La tercera particularidad es que la memoria no es lineal, sino que se asemeja más a una red jerárquica de eventos. Y la cuarta particularidad refiere a que las personas buscamos claves en nuestro contexto inmediato para recuperar la información almacenada en nuestra memoria.

Es claro que las personas algunas veces no pueden recordar un evento personal, aún cuando tienen a mano numerosas claves para recordarlo. Estudios sobre estos tópicos han encontrado que menos de una cuarta parte de los detalles críticos de eventos personales (identificados por las mismas personas como importantes) no pueden ser recordados un año después de sucedido el evento, y hasta un 50% de esos detalles no pueden ser recordados después de 5 años (Bradburn, Rips & Shevell, 1987). Existe evidencia de que la precisión del recuerdo decrece en pocas semanas y hasta en días para recuperar información

sobre actividades tan cotidianas como el consumo diario de alimentos (Smith, Jobe & Mingay, 1991, estudio 1).

A esto se le debe agregar que la recuperación de la información está afectada por las características de los eventos; por ejemplo, eventos raros (inusuales), recientes o frecuentes pueden ser más fácilmente recordados, mientras que las conductas rutinarias e irrelevantes son más difíciles de recordar. Tratemos de traer a la memoria por un momento cuántas horas pasamos sentados el día de ayer: ¿Qué responderíamos si nos hacen una encuesta sobre estos aspectos? (Tourangeau, et al., 2004).

También es claro que la memoria semántica y la memoria episódica poseen muchas características en común, pero también muchas diferencias (Tulving, 2002). En pacientes con daño prefrontal, por ejemplo, la memoria para recordar la fuente de la información se encuentra particularmente afectada, no así el conocimiento en sí mismo. Es decir, la información se recuerda correctamente, pero el contexto espacio-temporal en el que dicha información se adquirió no puede ser recordado (Tirapu-Ustárroz & Muñoz-Céspedes, 2005). Estudios en laboratorio indican también que la memoria episódica solo puede recuperar información que ha sido previamente almacenada, mientras que la memoria semántica tiene acceso a información que no ha sido explícitamente almacenada (Tulving, 2002). Pensemos por un momento en todas las cosas que sabemos, pero de las que no podemos precisar el momento y/o el lugar donde las aprendimos Estos estudios también indican que la información episódica es muy sensible a la interferencia y al olvido, porque depende de la condición temporal del evento (el momento en que lo vivimos). La memoria semántica, por su parte,

dispone de una información bien enraizada en una estructura relativamente estable de relaciones, al menos en las personas adultas, y en consecuencia es menos vulnerable a interferencias (Tulving, 2002).

La evidencia sugiere además que la memoria episódica está organizada en varios niveles diferentes de conocimiento personal, que se distinguen por su grado de generalidad (Conway & Pleydell-Pearce, 2000). En otras palabras, podemos pensar en la memoria autobiográfica como una red jerárquica compuesta en su nivel superior por períodos extendidos ("el tiempo que trabajabamos en la UCR"), en su nivel intermedio, por eventos duraderos y de carácter repetitivo ("realizando la investigación sobre EC") y en su nivel inferior, por eventos específicos e inusuales ("cuando sucedió el conflicto entre Costa Rica y Nicaragua por el dragado del Río San Juan").

Esta red posibilita recordar eventos del pasado a través de distintas vías: de arriba-abajo en la jerarquía, de manera secuencial a lo interno de los eventos duraderos ocurridos en períodos extendidos, y de manera paralela a partir de eventos inusuales (Schwarz y Oyserman, 2001).

Finalmente la investigación muestra que así como utilizamos pistas contextuales para comprender las preguntas de un cuestionario, así también buscamos pistas contextuales para recuperar la información que se nos solicita. Por ejemplo, es sabido que la información se recupera mejor cuando las claves utilizadas para recuperar la información estaban presentes a la hora de codificarla por primera vez, cuando el contexto en que la estamos recordando es similar al contexto de codificación y cuando nuestros estados psicológico y fisiológico también son similares al de ese momento (Jobe, 2003).

Pero como ya lo indicamos, en las entrevistas o aplicaciones de cuestionarios realizamos preguntas fuera del contexto original en el que ocurrieron, y esto plantea una serie de retos (o problemas) para la recuperación de la información. Dada la complejidad del proceso de recuperación de información, no es de sorprender que las personas pongan en marcha una serie de estrategias para aproximarse a una respuesta estimada, como veremos en las siguientes líneas.

Estimación de la respuesta

La valoración o estimación de la respuesta se refiere al uso de heurísticos o reglas de aproximación a la respuesta (Collins, 2003). En otras palabras, las personas estiman sus respuestas "a ojo de buen cubero". Detrás de estos heurísticos se encuentran las teorías implícitas que tienen las personas sobre sí mismas, sobre los otros y sobre el contexto, las reglas implícitas de interacción y la comunicación social.

Pensemos por un momento en una pregunta simple como "¿Cuántas horas trabajó usted la semana pasada?". La investigación sugiere que ante esta pregunta un entrevistado tiende a utilizar teorías implícitas sobre él o ella, su experiencia personal y su empleador, más que dar información sobre lo que sucedió la semana pasada.

Esto es, él o ella puede suponer que no hay razón alguna para asumir un cambio radical en su jornada laboral en este período de tiempo, dado que es una persona que regularmente va al trabajo y que su empleador cumple con las

regulaciones laborales básicas. Así, la persona en cuestión utilizaría las horas de trabajo estipuladas en su contrato laboral para estimar la respuesta final.

Igualmente la investigación empírica ha mostrado que las personas tendemos a "redondear" nuestras respuestas. Por ejemplo, reportes sobre lapsos ocurridos, ante preguntas del tipo "¿Hace cuantos días…?", "¿con qué frecuencia…" o "¿cada cuanto…?" evidencian agrupamientos de las respuestas en 7, 15 y 30, respectivamente, coincidiendo con la métrica usual (en Occidente) para el conteo de los días (Schwarz & Oysermna, 2001).

Finalmente, la investigación empírica ha mostrado que las personas sobreestimamos la frecuencia de eventos peculiares e inusuales y subestimamos la de eventos habituales (Schwarz y Oyserman, 2001).

Todas estas estrategias de aproximación a las respuestas dependen de la forma en que se planteen las preguntas en los cuestionarios, por tanto, deberían ser evaluadas antes de aplicar los cuestionarios en los estudios principales.

Ejecución de la repuesta

La ejecución de la respuesta hace referencia al acto mismo de expresar verbalmente, escribir o marcar una de las opciones de respuesta ante los reactivos o preguntas.

En este nivel, se ha observado que las personas "escanean" las opciones de respuesta para identificar aquellas que mejor se ajusten a la respuesta pensada; de hecho, se ha visto que el formato de respuesta afecta directamente la ejecución de esta. (Schwarz & Oyserman, 2001).

Piénsese en una pregunta como "¿Cuándo fue la última vez que usted escuchó hablar del referéndum sobre el Tratado de Libre Comercio de Estados Unidos con Centroamérica y el Caribe?", aunada a las siguientes opciones de respuesta: "a) Hoy, b) Ayer, c) Hace dos días, d) Hace una semana, e) El mes pasado". Este tipo de opciones de respuesta obliga a las personas a concentrarse en un período relativamente corto e induce a suponer que deben haber escuchado algo sobre el tema recientemente (Jobe, 2003).

A todos esto se le debe adjuntar los procesos psicofisiológicos que pueden influir en las respuestas, como la fatiga y el estado de ánimo a la hora de contestar el cuestionario; o bien las tendencias motivacionales básicas de los seres humanos, como la necesidad de ser aceptados, la necesidad de control y la deseabilidad social. Evidentemente los procesos cognitivos no están separados de los procesos motivacionales; todo lo contrario, la investigación sugiere que la emoción interactúa con la memoria, y la evidencia neurocientífica indica que la amígdala juega un papel muy importante en ello, particularmente en las fases de consolidación del recuerdo (Phelps, 2006).

La investigación muestra que la fase final de ejecución de la respuesta puede estar particularmente afectada por la tendencia natural en los seres humanos de "salvar la cara" y "quedar bien". Al respecto, la deseabilidad social es una de las tendencias más estudiadas y refiere a la tendencia a convenir con las normas sociales reportando con mayor facilidad conductas socialmente esperadas que conductas ilícitas y socialmente sancionadas (Collins, 2003; Jobe, 2003; Schawarz & Oyserman, 2001).

Es por todo esto que algunos autores suponen una quinta etapa de "edición" de la respuesta, para indicar que antes de marcar una respuesta, las personas están atentas a las normas sociales y a las reglas de interacción interpersonal, de modo que a partir de las expectativas de los otros (presentes o imaginados) revisan y corrigen sus respuestas.

Un ejemplo de nuestro laboratorio nos permite ilustrar este punto (Smith-Castro, Molina & Castelain, 2010). En una encuesta que realizamos entre 87 personas costarricenses entre los 16 y los 51 años de edad (63% mujeres), estudiamos el efecto de las demandas sociales relacionadas con la opinión sobre la unión civil entre personas del mismo sexo. Para ello confeccionamos dos tipos de cuestionarios: en una de las versiones, los participantes fueron primero interrogados sobre su aceptación o rechazo de la iniciativa sobre este tipo de unión civil y luego hicimos una serie de preguntas sobre prácticas religiosas (rezar, ir a la iglesia, leer la biblia), mientras que en la otra versión, los participantes contestaron primero las preguntas sobre religiosidad y luego brindaron su opinión sobre la unión civil entre personas del mismo sexo. Las dos versiones fueron distribuidas al azar entre los participantes. Los resultados de la consulta en función de las versiones del cuestionario se pueden observar en la Figura 2.

Aunque el reporte de prácticas religiosas no varió dependiendo de la versión del cuestionario, la opinión sobre este tipo de uniones civiles sí se alteró en función del orden de presentación de las preguntas (X^2 4.29, gl = 1, p = .039). Cuando les preguntamos directamente a las personas sobre su opinión sobre esta iniciativa, el porcentaje de apoyo fue del 67% y el rechazo, del 30%, mientras que

cuando se pusieron de relieve las prácticas religiosas antes de preguntar por la opinión, el acuerdo disminuyó al 45% y el rechazo aumento al 55%.

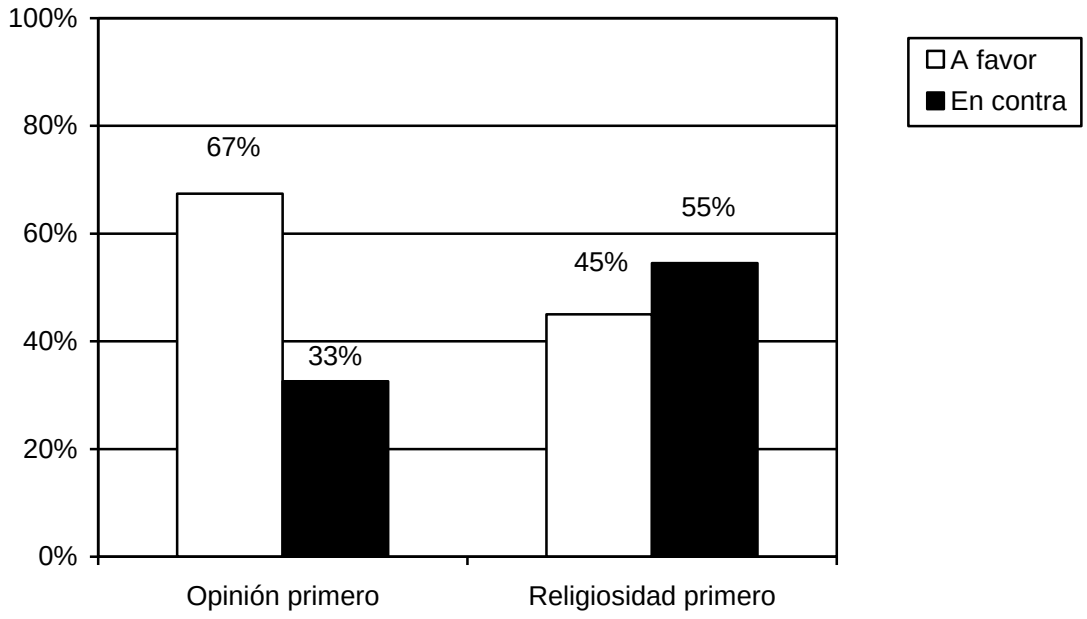


Figura 2. Apoyo a las uniones civiles gay según el orden de presentación de los reactivos

Nota. **Opinión primero**: En esta versión del cuestionario se solicitó la opinión sobre la unión civil antes de preguntar por las prácticas religiosas. **Religiosidad primero**: En esta versión se solicitó la información sobre las prácticas religiosas antes de preguntar por la opinión sobre la unión civil.

Nuestros resultados son consistentes con la idea de que activar la presencia de normas sociales como la religiosidad, directamente vinculadas con la temática en cuestión, pudo haber afectado la expresión del apoyo a esta iniciativa.

No solo el orden en que aparecen las preguntas en el cuestionario afecta la forma en que las personas responden. El tipo de aplicación del cuestionario es fundamental. Tourangeau & Smith (1996) encontraron que los cuestionarios

autoadministrados apoyados por el uso de computadoras aumentan la disposición de los encuestados a reportar conductas privadas y “embarazosas”, como la cantidad de compañeros sexuales o la frecuencia de uso de drogas ilícitas.

En síntesis, el modelo de cuatro etapas (comprensión-recuperación-estimación-ejecución) es una herramienta teórica que nos ayuda a considerar todos aquellos elementos cognitivos implicados en el proceso de contestar instrumentos de papel y lápiz, teniendo en cuenta siempre que, tal y como se desprende del modelo, recuperar eventos de la memoria autobiográfica resulta en sí un proceso complejo.

Las implicaciones para la construcción de instrumentos de papel y lápiz son claras: las preguntas o reactivos y el cuestionario en sí mismo deben ser acompañados de elementos que faciliten la comprensión, la recuperación de la información, la estimación de la respuesta y la ejecución de la misma. Evaluar si estos intentos son exitosos es la tarea de la EC, como veremos en la siguiente sección.

Literatura recomendada

De los trabajos que conocemos, el trabajo de Tourangeau, Rips & Rasinski (2004) es quizá la mejor recopilación sobre la psicología del proceso de respuesta de cuestionarios en el marco de encuestas. Otro nombre clave es Norbert Schwarz (Schawarz & Oyserman, 2001), cuyos trabajos citamos sistemáticamente en este apartado.

En el año 2003 la Revista *Quality of Life Research* dedicó su número 12 a los métodos cognitivos de evaluación de instrumentos de encuestas. De allí provienen los trabajos de Collins (2003) y Jobe (2003) que citamos aquí.

Finalmente, el Centro Nacional de Estadísticas en Salud de los Estados Unidos (NCHS), dedica una serie completa de investigaciones sobre cognición y medición en encuestas que pueden ser descargadas gratuitamente de la siguiente dirección electrónica: http://www.cdc.gov/nchs/products/series.htm

LA ENTREVISTA COGNITIVA EN LA PRÁCTICA

En las secciones anteriores tratamos de describir para qué se hace la EC y por qué es importante hacerla. En esta sección describimos qué es una EC y cómo se hace. Creemos que la mejor manera de transmitir cómo funciona esta técnica es mostrando de una vez la consigna con la cual se instruye a los participantes a la hora de realizarla (ver Cuadro 4).

Cuadro 4. Consigna para iniciar la entrevista cognitiva

Instrucciones
Estamos realizando una serie de pruebas para ver cómo funciona este cuestionario. Para eso yo le voy a dar el cuestionario y le voy a pedir que lo llene como si estuviéramos realizando la encuesta. En esta etapa lo que nos interesa es saber cómo está funcionando el cuestionario. Por eso le voy a pedir que conforme lo va completando ***piense en voz alta***. Es decir, que diga en voz alta todo lo que se le viene a la mente conforme va completando las preguntas. En cada pregunta, *yo le voy a realizar más preguntas sobre la redacción de estas*, las instrucciones y las opciones de respuesta. Es importante que tenga presente que nosotros queremos saber si el cuestionario funciona. No dude en decirme si algo le parece confuso y si algo se puede mejorar. Vamos a durar aproximadamente ____ minutos en todo el proceso. Antes de iniciar ¿tiene alguna pregunta? Práctica: Para irse acostumbrando a pensar en voz alta, vamos a practicar con la primera sección del cuestionario. A partir de la segunda sección vamos a iniciar formalmente con la entrevista.

Adaptada de Willis (2005).

Como se puede inferir de la consigna, la EC es una entrevista individual semiestructurada en ambiente controlado en la cual se le solicita contestar la

entrevista o completar el cuestionario (si éste es autoaplicado), y se le interroga sobre los distintos aspectos que intervienen o interfieren en el proceso de contestar las preguntas.

Se puede observar también que a la persona se le piden dos cosas básicas: decir en voz alta todo lo que está pensando mientras completa el cuestionario y contestar una serie de preguntas sobre el cuestionario. Estas son las dos técnicas básicas que se utilizan en la EC: pensar en voz alta y las tareas o pruebas verbales.

Pensar en voz alta

Cuando se utiliza esta técnica en el marco de una evaluación del funcionamiento de un cuestionario, el investigador o investigadora le pide a la persona participante que vaya completando el cuestionario o la entrevista y verbalice todo el proceso mental que le ocurre mientras llena el cuestionario (Erickson & Simon, 1993). El entrevistador o entrevistadora interviene lo menos posible mientras la persona va narrando todo lo que pasa por su cabeza y registra (por medio de audio y/0 video) el proceso mental verbalizado para detectar problemas a la hora de contestar los reactivos.

Las pocas intervenciones del entrevistador o entrevistadora se dirigen a pedirle a la persona que diga en voz alta lo que piensa en ese momento guiándose por la conducta verbal y no verbal del entrevistado o entrevistada (Presser, Couper, Lesser, Martin, Martir, et al., 2004). Un ejemplo de esta técnica se presenta en el Cuadro 5.

Cuadro 5. Ejemplo del método de “pensar en voz alta”

Entrevistador: Bueno (,) vamos a iniciar formalmente la entrevista (…) Entonces voy a pedirle que en voz alta la haga (…) Para practicar vamos a hacer esta parte de información sociodemográfica.

Participante 9: ¿Y lo voy llenando entonces?

Entrevistador : Sí

Participante 9: ¿Entonces tengo que leer lo que dice ahí y las respuestas que doy también tengo que decirlas?

Entrevistador: Por favor

Participante 9: OK. Nacionalidad (…) de Costa Rica. Sexo (…) femenino. Orientación sexual, homosexual, bisexual, heterosexual (*lee todas las opciones y enfatiza la respuesta en la tercera)*. Estado civil (,) soltera (…) Edad (,) 24 (…) Carrera que cursa (…) psicología y también comunicación colectiva (…) Nivel de la carrera en la que cursa la mayoría de las materias (,) sería psicología (,) entonces sería (,) nivel de la carrera (…) diay sería tercer año (,) pero me resulta confusa esa pregunta porque uno puede pensar en que uno está en la universidad o si está en tercer año o si está llevando todas las materias pero sería mi tercer año.

Aquí lo que debemos aclarar es la necesidad de “entrenar” al entrevistado o entrevistada en cómo llevar a cabo la tarea de pensar en voz alta. Se puede ayudar a la persona a verbalizar lo que pasa por su mente de distintas maneras. Una de las maneras más comunes es solicitar a la persona que se ubique mentalmente en la situación que se le está preguntando. Algo así como:

> *“Trate de visualizar su casa en este momento, empiece a caminar por su casa, por la sala, el comedor, la cocina, los cuartos, los baños… Ahora por favor dígame qué está viendo durante este recorrido. Ahora indíqueme que piensa mientras ve su casa”.*

Este tipo de entrenamiento le permite a la persona reconocer que debe verbalizar todo lo que se le viene a la mente mientras se ubica en “la situación” que se le consulta. Se necesitan varios ensayos de entrenamiento antes de iniciar formalmente con la entrevista, dependiendo de qué tan buena es la persona en verbalizar lo que pasa por su mente.

Willis (2005) describe dentro de las principales ventajas de esta técnica el hecho de que la producción de los participantes está poco contaminada por las expectativas del entrevistador, porque hay poca injerencia de parte de él o ella. Además, para realizarla se necesita de un entrenamiento mínimo, por lo que puede ser aplicada por varias personas al mismo tiempo y ahorrar costos. Finalmente la técnica permite observar in situ los problemas con que se enfrenta la persona a la hora de contestar el cuestionario conforme va completando el cuestionario.

La principal desventaja de la técnica es que depende de las capacidades de los y las entrevistados de verbalizar todo lo que pasa por su mente. Y aquí se han observado varios problemas (Willis, 2005). Por un lado se ha observado resistencia de parte de los entrevistados para expresar todo lo que se les ocurre y simplemente se limitan a contestar en voz alta, ya sea porque les parece irrelevante decir lo que piensan en ese momento, o bien porque temen decir lo que se les ocurre. Pensar en voz alta es una metacognición (pensar en qué se está pensando) y eso puede agotar a los entrevistados o bien puede influir en sus respuestas.

Pruebas cognitivas de reporte verbal

Las pruebas son preguntas directas que tienen como objetivo principal obtener evidencia sobre la manera en que los reactivos están siendo comprendidos, la medida en que los reactivos facilitan la recuperación de la información, el tipo de heurísticos que se utilizan para estimar las respuestas y las dificultades para hacer efectivas las respuestas.

Las tareas o pruebas de reporte verbal siguen los principios teóricos del modelo de cuatro etapas de Tourangeau (1984), pero podemos atribuirle a autores como deMaio, Mathiowetz, Rothgeb, Beach & Duran (1993) y a Willis (2005) el haber incorporado este conocimiento en el diseño de protocolos para evaluar potenciales problemas a la hora de contestar un cuestionario.

De esta literatura podemos distinguir al menos 5 tipos de pruebas que pueden realizarse para detectar problemas en el procesamiento de un cuestionario. Estas son: a) el parafraseo, b) los juicios de confianza, c) las pruebas de recuerdo, d) las pruebas de especificación y e) el sondeo del proceso mental (ver Cuadro 6).

Estas pruebas son aplicadas a los tres aspectos centrales de todo cuestionario, a saber: las introducciones en donde se especifican las instrucciones para contestar los reactivos, los reactivos en sí mismos y las opciones de respuestas que ofrecemos con cada reactivo. Finalmente estas deben tratar de detectar problemas en los cuatro momentos del proceso pregunta-respuesta, a saber, comprensión, recuperación de la información, estimación y ejecución (Willlis, 2005).

Cuadro 6. Ejemplos de tareas cognitivas de reporte verbal

Tipo de prueba	Ejemplo
Parafraseo	Podría decirme con sus propias palabras lo que acaba de leer
Juicios de confianza	¿Qué tan seguro está usted de esa respuesta?
Pruebas de recuerdo	¿Cómo llegó usted a recordar esa respuesta?
Pruebas de especificación	¿Podría decirme qué significa para usted la palabra…?
Sondeo del proceso mental	¿Qué hizo usted mentalmente para llegar a esa respuesta?

Traducido y adaptado de Willis (2005)

Tomando en cuenta todos estos elementos, nosotros hemos identificado 6 tareas centrales de las pruebas verbales. Otros autores incluyen más aspectos por evaluar (Willis, 2005), pero en nuestra experiencia estos son los fundamentales: a) evaluar las introducciones o instrucciones en términos de su comprensión; b) evaluar los reactivos en términos del significado que desean transmitir; c) considerar los supuestos o lógica subyacente a las preguntas; d) detectar los potenciales problemas para recordar la información solicitada; e) revisar si las preguntas o palabras que se utilizan en el cuestionario pueden herir susceptibilidades o pueden ser ajenas, extrañas u ofensivas; y f) verificar cuán adecuadas son las categorías de respuesta de cada pregunta en términos de a, b, c, d y e.

Así, si deseamos verificar si las personas comprenden el significado de las instrucciones podemos utilizar una prueba de parafraseo como “¿me puede decir con sus propias palabras que dicen las instrucciones?”.

En el caso de que queramos verificar el significado de los términos y conceptos utilizados en los reactivos podemos hacer pruebas de especificación del tipo "¿Qué significa para usted la palabra…?, ¿podría darme ejemplos?". También podemos guiarnos por la forma en que las personas van contestando el cuestionario y preguntar "usted dijo depende… ¿depende de qué?".

Para detectar problemas en la recuperación de la información tenemos varias posibilidades dentro de las pruebas de recuerdo ("¿Recuerda usted cuándo le sucedió algo así?") y sondeos del proceso mental ("¿Cómo llegó usted a esa respuesta?"). De nuevo, podemos guiarnos por la conducta del entrevistado: "Veo que arrugó la cara… ¿en qué estaba pensando en ese momento?".

Para verificar si nuestras preguntas están siendo ofensivas tenemos a disposición varias pruebas de especificación como "¿Cree usted que las personas se sientan cómodas con esta pregunta?", "¿siente usted que uno puede dar sinceramente una respuesta a esta pregunta?", "¿la palabra x le parece correcta?".

Y en el caso del análisis del proceso final de ejecución de la respuesta, se pueden utilizar las mismas pruebas aplicadas a las opciones de respuesta, como por ejemplo "¿Pudo encontrar en estas opciones la respuesta que usted quería dar?, "¿qué significa para usted ´totalmente en desacuerdo´ en esta lista de opciones?".

Como se puede observar existe una gran variedad de estrategias que se pueden utilizar para revisar el proceso de respuesta de un cuestionario. Una lista detallada de cada tarea de evaluación con algunos ejemplos de las pruebas cognitivas disponibles se presenta en el Cuadro 7.

A primera vista, puede ser muy abrumador tomar en cuenta tantos aspectos por evaluar y tantas pruebas para hacerlo. Algunos investigadores se ven tentados a (y a la vez desanimados por) realizar todas y cada una de las pruebas para los reactivos de sus cuestionarios. Esto implicaría una tarea impensable, porque por lo general los cuestionarios constan de varias secciones con varias escalas o índices que a su vez están compuestas por múltiples reactivos. De tal manera que nosotros recomendamos pensar en el Cuadro 7 como un catálogo de aspectos y pruebas que el investigador o investigadora ajusta a la medida de su cuestionario y sus objetivos de investigación.

Por ejemplo, si la investigación exige de las personas encuestadas un esfuerzo para recordar conductas pasadas, el investigador o investigadora podrá concentrarse solamente en explorar el aspecto 4, dedicado a detectar problemas de memoria y conocimiento. Si se trata de una encuesta con temas “delicados”, entonces la entrevista puede concentrarse principalmente en el análisis de sensibilidad.

Una ventaja obvia de pensar en el Cuadro 7 como un catálogo es su carácter logístico y económico. Lejos de hacer una o dos entrevistas largas para cubrir todos los puntos expuestos en este cuadro, el investigador o investigadora podrá realizar una mayor cantidad de entrevistas más cortas, focalizadas, estructuradas y económicas. El investigador o investigadora puede también escoger solo ciertos reactivos que le parezcan dudosos o ciertas escalas que no han sido probadas o validadas con anterioridad. De esta manera evita los problemas típicos de los procedimientos de recolección muy extensos como la fatiga, el aburrimiento, el aprendizaje, etc.

Asimismo, la gran ventaja de pensar en el cuadro 7 como un catálogo de aspectos y pruebas es de carácter científico. Para llevar a cabo la EC en el pilotaje del instrumento, el investigador o investigadora debió primero haber pensado muy bien cuál es el objetivo de su investigación, cuáles son los constructos centrales que desea estudiar, en qué medida sus instrumentos son operacionalizaciones idóneas para captar sus constructos y cuáles son las demandas que exige de los y las entrevistados. Es decir, antes de realizar la EC, este catálogo le ofrece la oportunidad al investigador de tomar distancia con respecto a su cuestionario, reflexionar si está bien planteado y eso lo devuelve a pensar sobre los aspectos centrales de su investigación.

Willis (2005) destaca dentro de las ventajas de esta técnica el hecho de que existe un mayor control y direccionalidad de la entrevista en comparación con la técnica de pensar en voz alta. Por otro lado, requiere poco entrenamiento por parte del sujeto, como sí se requiere cuando de pensar en voz alta se trata.

Dentro de las principales desventajas se encuentra la artificialidad, ya que introducir pruebas y sondeos puede producir una situación distinta a la que se sucede en el campo cuando se aplica el cuestionario definitivo. Finalmente existe el potencial problema de inducir sesgos en el entrevistado.

Nuestra experiencia indica que algunos entrevistados pueden sentir la necesidad de justificar sus respuestas en lugar de decir libremente lo que sucede en su mente, o bien que podemos provocar la tendencia en el sujeto de criticar indiscriminadamente cualquier palabra dentro del cuestionario, de tal manera que al final la entrevista se vuelve poco productiva.

Todas estas desventajas pueden controlarse haciendo un uso adecuado de la técnica de entrevista semiestructurada. Posteriormente ahondaremos un poco más sobre estos temas. Primero deseamos describir resumidamente otros tipos de pruebas que se pueden realizar en el marco de la EC.

Otras técnicas de sondeo

Además de estas pruebas verbales existen otras posibilidades de sondeo en el marco de la EC. Lo que describimos anteriormente se conoce como sondeo concurrente. En este tipo de sondeo las pruebas se van a realizando conforme la persona va completando el cuestionario. Pero existe también el sondeo retrospectivo, que se realiza una vez que la persona ha completado el cuestionario y es conocido también como *debriefing*, un anglicismo utilizado para describir una reunión posterior a una misión que se ha realizado y que en este caso constituye una valoración de lo realizado por parte del entrevistado (Willis, 2005).

Además existen pruebas que combinan el reporte verbal con la ejecución de tareas específicas. Dentro de estas tareas específicas destacan las tareas de clasificación de los reactivos de acuerdo con criterios teóricos, en la que se les pide a las personas entrevistadas, por ejemplo, clasificar los reactivos que ellas consideran difíciles en una categoría y los fáciles en otra; o bien, catalogar aquellas preguntas más cómodas de contestar en una categoría y las más incómodas en otra.

También se utilizan los puntajes de confianza, en donde se les solicita a los sujetos estimar en qué medida comprendieron los reactivos o están seguros de sus respuestas. A los entrevistados se les pueden presentar tarjetas con una

escala de calificación de la dificultad de los reactivos o de la seguridad con que emitieron sus respuestas.

Finalmente, se utiliza la medición de los tiempos de respuesta (mediante cronómetro), bajo el supuesto de que altas latencias de respuesta sugieren que los sujetos tienen problemas para contestar los reactivos (Collins, 2003; Jobe, 2003; Tourangeau, et al., 2004).

En síntesis el objetivo de la EC es realizar un sondeo minucioso sobre los potenciales problemas que tienen los entrevistados a la hora de contestar el cuestionario tomando en cuenta las cuatro etapas de Tourangeau (1984) del proceso pregunta-respuesta y considerando todos los elementos de un cuestionario, desde las instrucciones, pasando por cada una de las preguntas o reactivos, hasta las opciones de respuesta.

Este sondeo se lleva a cabo mediante la consulta directa a las personas entrevistadas mediante preguntas o tareas específicas para detectar problemas en cada etapa de proceso pregunta-respuesta y cada elemento del cuestionario En la siguiente sección presentamos algunos ejemplos de aplicación de esta técnica para evaluar cuestionarios autoaplicados.

Cuadro 7. Catálogo de pruebas cognitivas

Tarea	Problema	Pruebas
1. INSTRUCCIONES: Identifique problemas en todas las introducciones, instrucciones o explicaciones desde la perspectiva del encuestado/a.	Introducciones, instrucciones o explicaciones complicadas, confusas o vagas.	*¿Antes de pasar a las preguntas, podría repetirme con sus propias palabras la instrucción que acaba de oír o leer?*
2. CLARIDAD DE LAS PREGUNTAS O REACTIVOS: Identifique problemas relacionados con la intención o significado de las preguntas	La pregunta es muy larga o rara, la sintaxis es compleja, la redacción es incorrecta.	*¿Puede decirme con sus propias palabras que es lo que le acabo de preguntar?, ¿puede decirme con sus propias palabras la frase que acaba de leer?*
	Los términos técnicos están poco definidos, son muy complejos, no están claros	*¿Qué significa para usted la palabra (término) en esta pregunta?*
	Existen múltiples formas de interpretar la pregunta, la redacción es vaga o ambivalente.	*¿En qué estaba pensando usted cuando le pregunté sobre (tópico, término, etc.)?, ¿podría darme ejemplos?, ¿cuáles aspectos incluye el término (término)?, ¿cuáles no están contemplados en este término?*
	Los períodos de tiempo no están especificados, son vagos o contradictorios.	*¿Usted puede recordar ese período sobre el que le estoy preguntando?, Usted contestó (repetir la respuesta), ¿cuándo fue eso?, ¿qué período abarca eso?*
3. SUPUESTOS: Determine si existen problemas con los presupuestos o la lógica subyacente a las preguntas.	Los supuestos sobre la situación del entrevistado o su experiencia son inapropiados.	*¿Qué tanto se aplica esta pregunta a su experiencia?, ¿qué tan lejos está esta situación de su experiencia cotidiana?, ¿puede explicarme más su situación?*
	Se asume una experiencia o conducta constante y estable que en realidad puede variar.	*¿Diría usted que eso le pasa siempre?, ¿diría usted que eso varía dependiendo de algo?*
	El reactivo contiene más de una pregunta implícita o hace referencia a más de un objeto actitudinal.	*Usted acaba decir "depende", ¿depende de qué?; cuénteme más acerca de sus opiniones sobre este tema.*

Cuadro 7. (Continua)

Tarea	Problema	Pruebas
4. CONOCIMIENTO /MEMORIA: Verifique si los encuestados tienen problemas para saber o recordar la información que se solicita.	La persona no tiene conocimiento sobre el tema y es probable que no tenga una actitud al respecto.	*¿Qué tanto conoce sobre (tema)?, ¿qué tan seguro se siente usted al hablar sobre (tema)?*
	La actitud (evaluación) al respecto puede no estar consolidada o no existe.	*¿Había pensado usted antes en (tema)?, ¿qué tan frecuentemente ha pensado usted al respecto?*
	El entrevistado puede no recordar la información solicitada.	*¿Para usted fue esto fácil o difícil recordar?, ¿qué tan seguro está usted de eso?, ¿por qué?*
	La pregunta requiere de un proceso mental complejo.	*¿Qué fue lo que usted hizo mentalmente para contestar esta pregunta?, ¿en qué pensó primero?, ¿cómo llegó usted a esa respuesta?*
5. SENSIBILIDAD: Revise las preguntas o palabras de naturaleza sensible.	La pregunta hace referencia a tópicos privados, embarazosos, que implican conductas no deseadas o ilegales.	*¿Está bien hablar de estos temas en una encuesta o se siente muy incómodo?; en general, ¿cómo se siente usted ante este tipo de preguntas?*
	La redacción de los reactivos o los términos utilizados son poco sensibles, ofensivos o ajenos a la experiencia de los entrevistados.	*En esta pregunta, utilizamos el término (palabra sensible), ¿le suena bien a usted o utilizaría otro término?*
	El reactivo evoca respuestas socialmente aceptadas (deseabilidad social).	*¿Le parece que se puede dar cualquier respuesta a esta pregunta o más bien le parece que hay una respuesta correcta a esta pregunta?*
6. CATEGORÍAS DE RESPUESTA: Verifique qué tan adecuadas son las categorías de respuesta de cada pregunta.	Las preguntas abiertas son difíciles o inapropiadas.	*¿Fue fácil o difícil para usted decidir cuál respuesta dar a esta pregunta?*
	Existe desajuste entre la pregunta y las categorías de respuesta.	*En esta lista que le di, ¿fue fácil o difícil encontrar la respuesta que usted quería dar?*
	Los términos técnicos en las opciones de respuesta son poco claros, complejos o están sin definir.	*De esta lista, ¿qué significa para usted (término)?*

Cuadro 7. (Continua)

Tarea	Problema	*Pruebas*
	Las categorías de respuesta pueden ser interpretadas de múltiples maneras (vaguedad).	*Dígame qué se le viene a usted a la mente cuando le digo (categorías de respuesta) / cuando ve está escala de respuestas*
	Existe un traslape entre las categorías de respuesta.	*¿Qué tan fácil o difícil le resultó escoger la respuesta dentro de esta lista de opciones?, ¿por qué escogió usted esta respuesta y no las otras?*
	Hay opciones de respuesta ausentes.	*En esta lista que le di, ¿fue fácil o difícil encontrar la respuesta que usted quería dar?*
	El orden de las categorías de respuesta es ilógico.	*¿Fue fácil o difícil para usted decidir cuál respuesta dar a esta pregunta?; al ver esta lista de posibles respuestas, ¿usted encontró alguna dificultad para entenderla?*

Adaptado de Willis (2005).

Ejemplos de aplicación de la entrevista cognitiva

El primer ejemplo que deseamos ofrecer proviene de la tesis de licenciatura de la Licenciada Tatiana Aguiar (2009), quien tenía como objetivo adaptar una escala de neuroticismo para niños y niñas costarricenses de cuarto, quinto y sexto grado. La escala original, *Introversion und Neurotizismus bei Kindern* (INK), fue desarrollada en Alemania en los años 70 por Christel Nischan (1974, citada en Aguiar, 2009). Era necesario entonces someter la escala a un escrutinio minucioso mediante la EC antes de proceder a realizar los análisis psicométricos y adaptar la escala a nuestro contexto. Algunos ejemplos de los reactivos se pueden apreciar en el Cuadro 8.

Cuadro 8. Algunos reactivos de la escala de neuroticismo INK para niños y niñas.

1. ¿A veces tu corazón palpita duro?	Sí	No
2. ¿A menudo te quedas absorto en pensamientos?	Sí	No
3. ¿Te sales fácilmente de tus casillas?	Sí	No
4. ¿A menudo te quedas despierto tendido en tu cama?	Sí	No

Fuente: Aguiar (2009).

La autora entrevistó a 3 niños y 3 niñas de cuarto, quinto y sexto grado (una pareja por nivel) mediante EC. Algunos extractos de las entrevistas se pueden observar en el Cuadro 9.

Cuadro 9. Extractos de Entrevistas Cognitivas con niños y niñas costarricenses.

INSTRUCCIONES
E: Antes de iniciar contestando las preguntas, ¿me podrías decir con tus propias palabras las instrucciones que acabas de leer?
S: Tengo que hacer una equis para responder a estas preguntas, y tengo que pensar cómo soy casi todos los días… pero lo que leí está un poco largo.
CLARIDAD DE LOS REACTIVOS
E: ¿Me podrías contar con tus palabras qué es lo que se pregunta en la oración “A menudo te quedas tendido en tu cama”?
S: Me están preguntando si me quedo acostada en la cama antes de dormirme… se entiende mejor si uno dice ‘acostado’ en lugar de ‘tendido’…
E: ¿Qué te parece la palabra “casillas” en la pregunta “Te sales fácilmente de tus casillas”?
S: Es como que uno se enoja… pero yo lo sé porque eso dice mi mamá cuando está enojada. No sé si todas las personas saben qué son ‘casillas’…

Fuente: Aguiar (2009).

Debido a la corta edad de la población meta, la autora se concentró principalmente en detectar problemas de comprensión de los reactivos y en

adaptar el vocabulario al nivel de los y las encuestados. De los extractos se observa que los niños y niñas son capaces de indicar a la investigadora dónde ellos(os) mismos(as) y otros(as) de su misma edad podrían tener problemas a la hora de contestar los reactivos. Se observa también que requirieron de poco entrenamiento para saber que su tarea consiste en ayudar a mejorar la comprensión del cuestionario.

La autora reporta que la mayor cantidad de cambios en la escala original de neuroticismo se dieron en el nivel de la comprensión de los conceptos. Ejemplos de conceptos que debieron ser sustituidos son "resentimiento", "estar tendido", "salirse de las casillas", "sentirse miserable", "destrozar" y "hartarse" (Aguiar, 2009). Estos fueron entendidos por la mayoría de participantes, pero ellos mismos consideraron que eran inapropiados para niños y niñas de su edad.

En el caso de los términos "colérico" y "te acomplejas", la autora encontró que la comprensión era insuficiente, por lo que necesariamente debieron ser cambiados por "muy enojado" y "te avergüenzas". Asimismo, la autora notó que "sensibilidad" era un concepto comprendido únicamente por los y las estudiantes de sexto grado.

Pese a los cambios que fueron necesarios en la escala, la autora destaca que los niños y niñas poseían una amplia comprensión de sentimientos y emociones como alegría, enojo, soledad, culpa y vergüenza, al igual que los conceptos de amistad, cansancio, timidez, preocupación, ridículo y nerviosismo (Aguiar, 2009).

El segundo ejemplo que ofrecemos proviene de la tesis de la Licenciada Karla Ugalde (2009), quien tenía como objetivo estudiar el grado de estrés que

viven las personas refugiadas colombianas desde que salen de sus hogares en Colombia hasta que se asientan en Costa Rica. Para ello, la autora diseñó una escala en donde consultaba directamente a las personas refugiadas sobre el grado de estrés que experimentaron en distintos momentos de su proceso migratorio desde la decisión de salir de Colombia y tener que separarse de su familia, pasando por los trámites para solicitar refugio, buscar trabajo y conseguir vivienda, hasta los retos de adaptarse al modo de vida costarricense. Para cada fase del proceso de aculturación, los y las entrevistadas tenían a su disposición una escala de 1 (nada estresante) a 6 (muy estresante). El Cuadro 10 presenta algunos de los 33 reactivos que componen la escala.

Cuadro 10. Algunos reactivos de una escala de estrés percibido por aculturación.

¿Qué tan estresante (tenso o angustiante) le resultó…?	**Nada Estresante**				**Muy estresante**
Tomar la decisión de salir de Colombia.	1	2	3	4	5
Prepararse para salir de Colombia.	1	2	3	4	5
Esperar el resultado de la solicitud de refugio.	1	2	3	4	5
Conseguir trabajo.	1	2	3	4	5
Hacer amigos.	1	2	3	4	5
La forma de ser del costarricense.	1	2	3	4	5
Las costumbres costarricenses.	1	2	3	4	5

Fuente: Ugalde (2009).

La autora estaba particularmente preocupada por dos aspectos: las diferencias culturales en el uso de conceptos y el hecho de que hablar sobre el tema de refugio puede ser muy doloroso para los y las participantes. Ella entrevistó a cuatro personas refugiadas (2 hombres y 2 mujeres) mediante EC

para explorar estos aspectos. Dos extractos de esas entrevistas se presentan en el Cuadro 11.

Uno de los principales aprendizajes de estas entrevistas fue reconocer que aquellos conceptos considerados como comunes en nuestra cultura pueden resultar extraños para personas de culturas aparentemente muy similares a la nuestra. Conceptos cotidianos en Costa Rica resultaron ser ofensivos para las personas refugiadas. Adicionalmente, se modificó el cuestionario de tal manera que temas sensibles como la decisión de abandonar el país de origen (muchas veces asociado con eventos traumáticos) o altamente estresantes (como tener que esperar el resultado de la solicitud de refugio o los preparativos de la salida del país de origen) fueron introducidos en el cuestionario paulatinamente, de tal manera que fueran más fácilmente procesados por los entrevistados (Ugalde, 2009).

Cuadro 11. Extractos de Entrevistas Cognitivas con personas refugiadas.

COMPRENSIÓN
E: ¿Y qué entiende por palabras como bar, soda, restaurante? ¿Son palabras comunes?
S: Sí. Ahorita sí. Pero para una persona recién llegada, no. Son palabras. Un bar allá es donde solo mantienen las prostitutas. Entonces a mí me dicen, vamos a ir a este bar, al bar que queda allí (…) Este qué se está creyendo.
E: ¿Qué palabras serían adecuadas en este caso, digamos?
S: *Si es para colombianos, usted tiene que decir una fuente de soda o una discoteque.*
CONTENIDO SENSIBLE
E: ¿Y cómo se siente ante este tipo de preguntas?
S: "Pues, a veces sí es maluco, porque recuerda uno cosas que no quisiera volver a pasar por la mente. Sí por ese lado, sí se pone maluco. Ah, uno recuerda lo que le tocó pasar (…)"

Fuente: Ugalde, 2009.

El tercer ejemplo proviene de nuestro propio laboratorio (Smith-Castro et al., 2009). En este proyecto teníamos como objetivo someter una escala de medición de actitudes a distintos procedimientos de evaluación y mejoramiento. Específicamente queríamos comparar los cambios que se sugieren mediante EC con los cambios que se derivan a partir de un panel de expertos. En la siguiente sección describimos los resultados de la comparación de estas dos técnicas. Por ahora nos interesa dar un par de ejemplos de cómo se dieron las entrevistas para este caso particular de medición de actitudes.

Nuestro primer paso fue decidir el constructo o temática social por ser investigada. Decidimos trabajar sobre el tema de la homofobia, entendida como la animadversión o rechazo de la homosexualidad y a los homosexuales, debido a que se trata de una temática de gran relevancia social, especialmente a la luz de las últimas discusiones en torno al proyecto de ley sobre el matrimonio entre personas del mismo sexo.

Una vez seleccionado el constructo solicitamos a 6 estudiantes avanzados de psicología construir reactivos siguiendo una definición básica de homofobia como la anteriormente descrita. Los y las estudiantes poseían conocimientos básicos en construcción de escalas, pero no necesariamente sabían mucho sobre el constructo en cuestión. De esta manera recreamos una situación común en la investigación social: la necesidad de diseñar un cuestionario sin mucho conocimiento sobre el constructo y poca revisión previa de la bibliografía pertinente.

Los y las estudiantes construyeron 42 reactivos sobre la base de esta definición. A partir de estos reactivos construimos un instrumento experimental (ver Cuadro 12).

Cuadro 12. Algunos reactivos de una escala experimental de homofobia.

	TD	D	I	A	TA
1. Considero que la homosexualidad es una enfermedad mental o un problema, que tiene que ser tratado.	1	2	3	4	5
2. Tengo amigos (as) homosexuales.	1	2	3	4	5
3. ¿Considera a un hombre o mujer homosexual como una persona que es parte de una "moda" de la nueva generación?	1	2	3	4	5
4. Una persona que ha sufrido algún tipo de abuso sexual durante su infancia tiene altas probabilidades de convertirse en homosexual.	1	2	3	4	5
5. Me sentiría incómodo(a) si tuviera que trabajar de cerca con una persona homosexual.	1	2	3	4	5
6. Me sentiría desilusionado(a) si tuviera un hijo(a) homosexual	1	2	3	4	5
7. ¿Considera a un hombre o mujer homosexual como una persona inferior Enferma / cree que se puede curar?	1	2	3	4	5
8. Ante un hecho contra los homosexuales actuó con silencio	1	2	3	4	5
9. ¿Conoce familiares o amigos homosexuales?	1	2	3	4	5
10. Considero que las personas nacen con una predisposición a la homosexualidad.	1	2	3	4	5

Nota. TD = totalmente en descuerdo, D = en desacuerdo, I = indeciso(a), A = de acuerdo, TA = totalmente de acuerdo.

Este instrumento experimental fue evaluado mediante EC con 10 estudiantes universitarios. Cada entrevista tuvo una duración de una hora. Algunos extractos de las entrevistas se observan en el Cuadro 13.

Durante el proceso detectamos los siguientes problemas: a) dificultades en la redacción de los reactivos, b) vaguedad en la información que contienen, c) presencia de términos técnicos de difícil comprensión, d) reactivos de contenido sensible, e) activación de tendencias a la deseabilidad social, f) supuestos inadecuados sobre el conocimiento de los entrevistados, g) exigencia de procesos mentales complejos, h) presencia de más de un objeto

actitudinal en un mismo reactivo, i) opciones de respuesta faltantes y j) desajustes entre pregunta y categorías de respuesta.

Cuadro 13. EC con estudiantes universitarios.

CONTENIDO SENSIBLE
E: Una persona que ha sufrido algún tipo de abuso sexual durante su infancia tiene altas probabilidades de convertirse en homosexual" ¿Se siente incómoda al leer o contestar esa pregunta?
S: Diay me impacto porque no la vi llegar (,) pero SI incomoda de contestarla (,) no sentí que no me fueran a preguntar tal vez como dentro de las últimas preguntas tal vez si pero (…)
AMBIGÜEDAD
E: Queda claro, que hay dos cosas, una enfermedad mental y un problema. Eso remite a dos cosas distintas
S: Si yo siento que remite a dos cosas distintas, porque si hablamos de enfermedad mental hablamos de una persona enferma y yo siento que estaría en desacuerdo (,) pero si hablamos de problema (,) si se puede decir que para esa persona sea un problema que son de las personas que dicen que dicen si soy homosexual y me siento mal por eso (…)

Fuente: Smith-Castro et al. (2009).

Como se puede observar, la EC nos permitió detectar gran cantidad de problemas en un instrumento de medición de una temática altamente sensible y, como veremos en la siguiente sección, la aplicación de la EC permitió mejorar sustantivamente las propiedades psicométricas de la escala original.

Creemos que estos pocos ejemplos permiten ilustrar la utilidad de la técnica para detectar problemas en el proceso respuesta de cuestionarios autoaplicados. También creemos que dan una idea de su potencial uso para evaluar otros tipos de instrumentos, como entrevistas estandarizadas y guías de entrevista de grupos focales. Se trata en todo caso de pedir a las personas que potencialmente utilizarían el cuestionario que nos ayuden a detectar

problemas a la hora de contestar las preguntas que hacemos. En las siguientes secciones compartimos algunas recomendaciones sobre los aspectos prácticos que se deben considerar a la hora de utilizar esta técnica.

Aspectos operativos y logísticos

Como en cualquier entrevista, los aspectos claves de la EC son el entrevistador, el entrevistado y el contexto en que se lleva a cabo la entrevista.

El entrevistador requiere de un entrenamiento básico en entrevistas semiestructuradas. Toda EC tiene los componentes de cualquier entrevista cualitativa, a saber: a) presentación e introducción, b) calentamiento, c) desarrollo de los temas centrales y d) cierre. Y como en toda entrevista, se requiere que los entrevistadores posean la habilidad de crear rápidamente un ambiente positivo y una dinámica de entrevista agradable, relajada y de mutuo respeto (un buen *rapport*); de enfatizar en la perspectiva de la persona entrevistada; de adaptarse a las particularidades de la persona entrevistada; y de hacer buenas preguntas, es decir, de realizar preguntas directas, libres de sesgos y solicitando siempre elaboraciones, argumentaciones y ejemplificaciones de las respuestas (Flik, Kardoff, Jeupp, Rosentiel & Wolff, 1995; Mack, Moqueen, Guest & Namey, 2005; Mayring, 1993; Wengraf, 2006). Las recomendaciones básicas para todo entrevistador se listan en el Cuadro 14.

La mayoría de autores recomiendan llevar a cabo entrevistas individuales en lugar de entrevistas grupales o grupos de discusión, debido a que los procesos colectivos pueden afectar las respuestas de las personas. Además, los grupos focales se concentran en los consentidos socialmente compartidos, mientras que la EC gira alrededor del proceso mental individual

(Willis, 2005). Esto no quiere decir que no se puedan realizar entrevistas grupales para definir los contenidos de un cuestionario o detectar problemas generales de comprensión de los instrumentos. En realidad, son técnicas complementarias: la EC está es idónea para estudiar el proceso mental de los sujetos, los grupos de discusión, para indagar sobre los contenidos socialmente compartidos que se transmiten en el cuestionario.

Cuadro 14. Recomendaciones para los y las entrevistadores(as).

Poseer un buen dominio sobre el objetivo de la entrevista.
Asegurar que las condiciones (lugar, ambiente y equipo) sean las apropiadas.
Cerciorarse de que su apariencia sea la apropiada para la ocasión.
Elegir comentarios o preguntas de apertura correctas.
Acercarse paulatinamente a las preguntas centrales.
Monitorear el impacto de sus conductas sobre el entrevistado.
Evitar preguntas sesgadas (que inducen a respuestas cerradas).
Promover la elaboración del entrevistado.
Demostrar que se está escuchando.
Verificar lo que se comprende y solicitar aclaraciones de lo que no se comprende.

Fuentes: Flik, et al. (1995), Mack et al. (2005), Mayring (1993) y Wengraf (2006).

En lo que respecta a los entrevistados, la discusión actual se centra precisamente en decidir cuántos participantes reclutar para cada evaluación y cómo escogerlos (Castellano-Ackerman & Blair, 2006).

En los estudios que hemos revisado, normalmente se llevan a cabo como mínimo 5 y como máximo 20 entrevistas cognitivas para la evaluación de un instrumento (Willis, 2005), aunque existen estudios en donde se realizan

más de 90 EC. Lo importante aquí es la selección de los participantes, quienes, como es usual en todo muestreo cualitativo, son seleccionados intencionalmente.

Regularmente se sugiere hacer una selección por cuotas basado en las características de la muestra por consultar en el estudio principal (edad, sexo, procedencia geográfica, nivel educativo, etc.); o bien buscar a las personas siguiendo el principios de minimizar las diferencias entre los casos, con el fin de obtener las tendencias básicas (detectar la presencia de un determinado problema), 0 el de maximizar las diferencias entre los casos, con la intención de incrementar el rango de problemas por detectar.

En otros casos se utiliza el criterio de reclutar personas con características particulares, dependiendo de los objetivos del estudio (por ejemplo, fumadores para una encuesta sobre el uso del tabaco), pero también personas que no poseen la característica buscada (como control). Otros autores utilizan los principios básicos de la cognición humana como criterio para seleccionar a los entrevistados. Así, podemos pensar en reclutar para la EC dos o tres sujetos que formen parte del grupo de menor edad y dos o tres del grupo de mayor edad, dependiendo del rango de edad esperado en la encuesta principal (dos adolescentes y dos adultos mayores, por ejemplo), a sabiendas de que la memoria a corto plazo puede variar entre estos rangos de edad. O bien se pueden incluir personas con mayor o menor experiencia en el tópico investigado para tener un rango de todas las posibles condiciones de repuestas (Castellano-Ackerman & Blair, 2006).

En síntesis, el muestreo probabilístico no es un método que se use normalmente en los laboratorios de EC. Se usan principalmente muestras por

cuotas, tratando de obtener un rango variado de edades, sexos, niveles socioeconómicos, habilidades cognitivas u otras características relevantes para la investigación. En general, la selección de sujetos se hace dentro de los potenciales encuestados en el estudio principal, es decir, la EC no se realiza con jueces expertos. Los potenciales encuestados son los expertos que buscamos aquí.

En lo que respecta a las condiciones de trabajo, es claro que la EC necesita ser realizada en un ambiente controlado, libre de potenciales fuentes de interrupción y distracción. Los laboratorios de las grandes casas encuestadoras normalmente tienen facilidades como oficinas aisladas con cámaras de video, sistemas de grabación y posibilidades de observación remota (cámara de Gesell). Sin embargo, para llevar a cabo una buena EC no es necesario todo este equipo. Una oficina o un aula y hasta el comedor de una casa pueden servir de laboratorio temporal. Lo importante es asegurar las condiciones de privacidad necesarias para llevar a cabo la entrevista. El equipo mínimo que se requiere es una grabadora de audio, el protocolo con las pruebas cognitivas que se desean realizar, un diario de campo y el cuestionario de prueba (Willis, 2005).

Aunque las entrevistas cognitivas pueden llegar a durar hasta dos horas, la experiencia dicta que lo óptimo son entrevistas de una hora como máximo, y por lo general es necesaria una sola sesión por persona. Ahora bien, preparar una sola sesión de entrevista de aproximadamente una hora requiere un tiempo de preparación considerable, por lo cual se recomienda que una sola persona no lleve a cabo más de tres entrevistas al día.

En general, las entrevistas producen mucho material verbal que necesita ser procesado antes de tomar una decisión sobre los cambios por realizar en el cuestionario. A continuación presentamos algunas sugerencias para el procesamiento de la información, a partir de nuestra propia experiencia.

Estrategias de protocolización y análisis

En una EC, el material básico que debe ser procesado es el reporte verbal generado por la técnica de pensar en voz alta y las respuestas a las pruebas verbales. Pero también se pueden incluir todos los comentarios al margen que pueda dar el entrevistado y que sean relevantes para la entrevista (Ericsson & Simon, 1993).

El material puede ser procesado de muchas maneras y al parecer no existe una única forma de hacerlo (Willis, 2005). Algunas organizaciones instruyen a sus entrevistadores a escuchar las cintas de las entrevistas y anotar los problemas encontrados en cada reactivo evaluado, mientras que otras trabajan con las notas del diario de campo que se llevó durante la entrevista. Otras organizaciones piden una transcripción de las entrevistas.

En lo que respecta a la transcripción, existen también muchos métodos. Nosotros utilizamos la transcripción literal de las entrevistas utilizando códigos muy sencillos, recomendados por Maynring (1993) para tal efecto y que presentamos en el Cuadro 15.

Prácticamente la codificación implica poner entre paréntesis todo aquello que no fue verbalizado explícitamente (risas, enojo, irritaciones, énfasis, aumento del volumen de la voz, pausas, etc.). En el marco de nuestra investigación sobre el impacto de la EC en un instrumento de medición de homofobia (Smith-Castro, et al., 2009), la transcripción de las entrevistas

produjo textos de aproximadamente 8.000 palabras. Se trataba de entrevistas de una hora aproximadamente. Un extracto de una entrevista transcrita se presenta en el Cuadro 16.

Cuadro 15. Códigos de transcripción.

Símbolo	Significado
(,)	pausa muy corta (respiración)
(...)	pausa mediana
(pausa)	pausa larga (más de un minuto)
(?)	Pregunta
(subrayar el texto)	entonación en una palabra o frase, palabras en donde se pone un especial énfasis con la voz
(Mhm)	Señal de recepción, expresión para llenar una pausa
(risa)	expresiones no verbales en paréntesis
(voz baja)	expresiones no verbales en paréntesis
(voz alta)	expresiones no verbales en paréntesis
(enojo)	expresiones no verbales en paréntesis
(Hablar simultáneamente)	se anota entre paréntesis que hablan al mismo tiempo
(xxxxx)	Incomprensible

Fuete: Maynring (1993).

El análisis de los datos en el marco de una EC se concentra en detectar los potenciales problemas en las etapas del proceso de pregunta respuesta, para lo cual se pueden crear categorías para cada potencial problema presente. En nuestro estudio construimos 16 diferentes códigos o problemas (ver Cuadro 17); posteriormente identificamos la presencia de estos en cada entrevista. Para ello utilizamos el programa Atlas Ti, el cual nos permitió identificar aquellas producciones verbales referidas a cada una de las

categorías de análisis (ver Figura 3), calcular la frecuencia con que cada problema se presentaba en la muestra de 10 participantes y determinar cuáles sujetos reportaron el problema.

Cuadro 16. Ejemplo de transcripción de una entrevista.

E: Muy bien (,) antes de pasar a las preguntas (,) (¿) podría repetirme con sus propias palabras la instrucción que usted acaba de leer (?)

S9: Bueno (,) lo que dice es que tengo que contestar este cuestionario conforme a lo que es una persona homosexual y que bueno en una escala del 1 al 5 en donde uno está más de acuerdo entre varios puntajes (,) o sea 5 es como más de acuerdo y 1 es como menos de acuerdo

E: Muy bien (,) (¿) Entonces le pareció que la instrucción está clara (?)

S9: Está clara (,) pero está un poco larga

E: Un poco larga

S9: Da pereza leerla (,) Yo creo que no hace falta repetirlo tanto porque igual la gente aunque no lo hubiera entendido (,) solo hubiera entendido mejor solo viendo como la escala aquí

E: Muy bien. (¿)Qué significa para usted la palabra homosexualidad (?)

S9: Homosexualidad (,) significa una persona que le gustan las personas del mismo sexo(...) O sea (,) que a los hombres le gustan los hombres y a las mujeres le gustan las mujeres y que tienen relaciones sexuales digamos

E: Muy bien (,) (¿) existen en este párrafo palabras que son ambiguas, que tienen más de un sentido de comprender (?)

E: Muy bien (,) Entonces vamos a pasar a los reactivos siguientes (…) Entonces le voy a pedir por favor (…)

S9: Los leo (¿?)

E: sí

Fuente: Smith-Castro, et al. (2009).

El análisis final implica la interpretación de los problemas y la toma de decisiones sobre los cambios que deben realizarse en los reactivos. En la práctica se acostumbra a reunir a los investigadores y entrevistadores para tomar decisiones sobre cada reactivo. También pueden trabajar independientemente en la calificación de los textos y posteriormente calcular el acuerdo de la calificación mediante medidas de acuerdo como el Kappa de Cohen, las cuales se utilizan normalmente para la estimación de la objetividad de la calificación, como vimos en la primera sección de este Cuaderno (Neuendorf, 2002).

Cuadro 17. Potenciales categorías de análisis para una EC.

Categorías de análisis:
1. Vaguedad y contradicción
2. Términos técnicos difíciles (comprensión)
3. Períodos de referencia inadecuados
4. Supuestos inadecuados o ilógicos
5. *Double barreled* (dos objetos actitudinales en una sola frase)
6. Desconocimiento
7. Problemas de memoria
8. Actitud no formada
9. Proceso mental complejo
10. Contenido sensible
11. Deseabilidad social
12. Categorías de respuesta confusas
13. Categorías de repuesta inadecuadas
14. Incongruencia entre el reactivo y sus categorías de respuesta
15. Traslape en las categorías de respuesta
16. Opciones de respuestas ausente

Fuente: Smith-Castro, et al. (2009).

En nuestro caso sistematizamos la información en una tabla de Excel tal y como se presenta en la Figura 4 y en varias sesiones de discusión tomamos las decisiones sobre los cambios en los reactivos. En esta tabla incluimos el tipo de problema que se observó, la frecuencia con que se presentó, los entrevistados que lo detectaron, el cambio que sugerimos y el reactivo tal y como quedó transformado al final del proceso.

A modo de resumen, podemos decir que el procesamiento de una EC no difiere en mucho del procesamiento que normalmente se hace de otros tipos de entrevista y en general de todo tipo de datos cualitativos. En la práctica cotidiana y por razones de costo y tiempo, no siempre es posible realizar todo el proceso de transcripción, codificación y análisis que nosotros llevamos a cabo, sin embargo, es recomendable documentar lo más fielmente posible el proceso para la toma de decisiones.

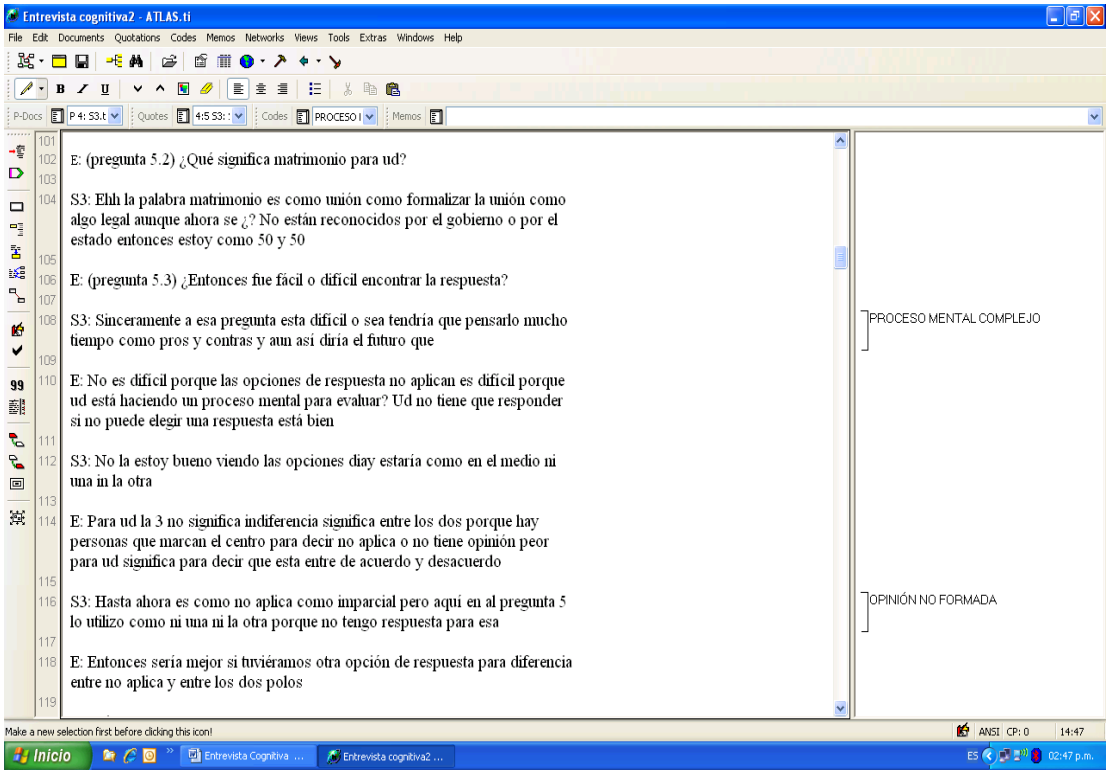
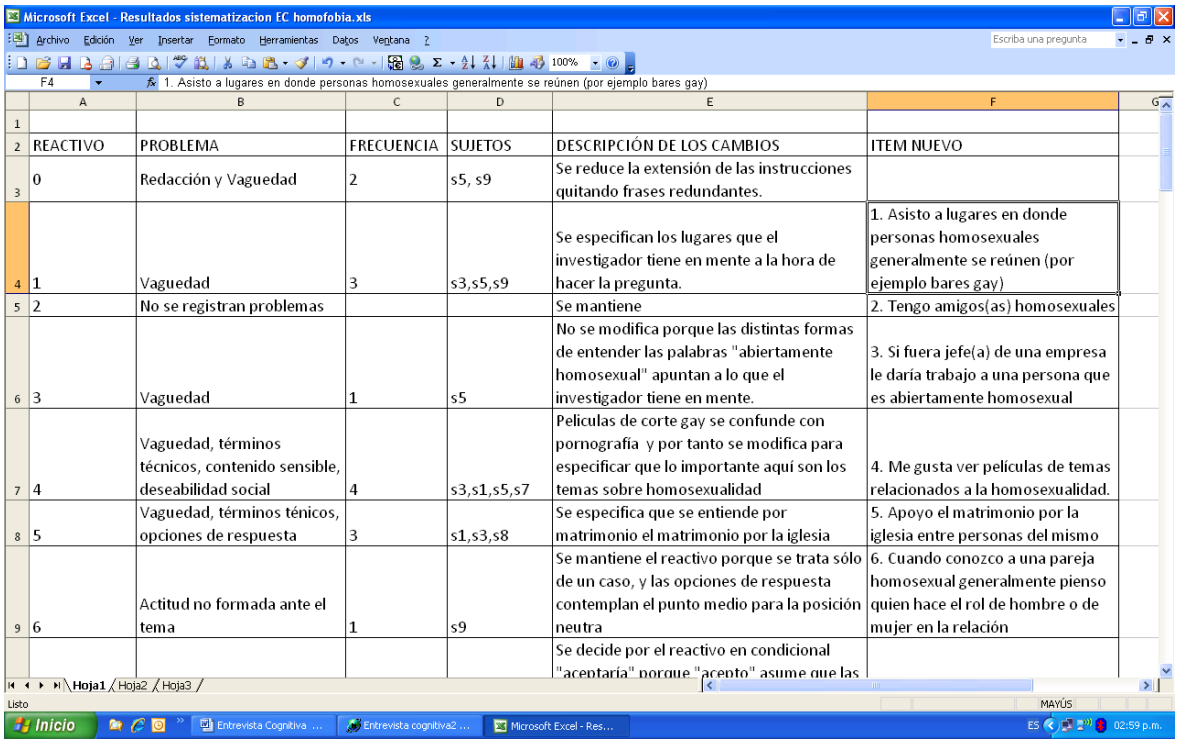


Figura 3. Procesamiento de las entrevistas en Atlas Ti.



REACTIVO	PROBLEMA	FRECUENCIA	SUJETOS	DESCRIPCIÓN DE LOS CAMBIOS	ITEM NUEVO
0	Redacción y Vaguedad	2	s5, s9	Se reduce la extensión de las instrucciones quitando frases redundantes.	
1	Vaguedad	3	s3,s5,s9	Se especifican los lugares que el investigador tiene en mente a la hora de hacer la pregunta.	1. Asisto a lugares en donde personas homosexuales generalmente se reúnen (por ejemplo bares gay)
2	No se registran problemas			Se mantiene	2. Tengo amigos(as) homosexuales
3	Vaguedad	1	s5	No se modifica porque las distintas formas de entender las palabras "abiertamente homosexual" apuntan a lo que el investigador tiene en mente.	3. Si fuera jefe(a) de una empresa le daría trabajo a una persona que es abiertamente homosexual
4	Vaguedad, términos técnicos, contenido sensible, deseabilidad social	4	s3,s1,s5,s7	Peliculas de corte gay se confunde con pornografía y por tanto se modifica para especificar que lo importante aquí son los temas sobre homosexualidad	4. Me gusta ver películas de temas relacionados a la homosexualidad.
5	Vaguedad, términos ténicos, opciones de respuesta	3	s1,s3,s8	Se especifica que se entiende por matrimonio el matrimonio por la iglesia	5. Apoyo el matrimonio por la iglesia entre personas del mismo
6	Actitud no formada ante el tema	1	s9	Se mantiene el reactivo porque se trata sólo de un caso, y las opciones de respuesta contemplan el punto medio para la posición neutra	6. Cuando conozco a una pareja homosexual generalmente pienso quien hace el rol de hombre o de mujer en la relación
				Se decide por el reactivo en condicional "aceptaría" porque "acepto" asume que las	

Figura 4. Sistematización de los datos en Excel

Literatura recomendada

Como se puede observar, esta sección se ha fundamentado principalmente en el trabajo de Gordon Willis (2005), quien es el autor que más ha trabajado en el tema de los protocolos de EC. Sin embargo, los autores más citados en lo que respecta a la técnica de pensar en voz alta y su respectiva protocolización son Ericsson & Simon (1993).

Tom Wengraf (2006) ha producido un texto dedicado exclusivamente a la entrevista cualitativa, desde su preparación hasta las estrategias de análisis, el cual puede ser muy útil para quienes desean profundizar en el tema.

Para las personas interesadas en las estrategias de análisis de contenido, recomendamos el texto de Neuendorf (2002). Por su parte, Lewins y

Silver (2007) ofrecen una descripción comparativa de distintos programas de cómputo que pueden ayudar a sistematizar y analizar los datos cualitativos. Para quienes están interesados específicamente en el uso del Atlas Ti, recomendamos el manual elaborado por la empresa creadora, el cual puede ser descargado del siguiente sitio web:

http://www.atlasti.com/uploads/media/atlman_01.pdf

¿SE TRATA DE UNA ESTRATEGIA EFECTIVA?

Existen al menos tres fuentes de escepticismo en torno a la aplicación de la EC en el marco de las encuestas. Algunos autores se preguntan si los sujetos son realmente capaces de reportar los procesos mentales involucrados a la hora de hacer efectivas sus respuestas. Otros autores se preguntan si estos métodos difieren sustancialmente de los enfoques más tradicionales, como un buen panel de jueces expertos o el ojo entrenado de un buen entrevistador. Finalmente, hay quienes dudan del impacto real de los cambios producidos en el laboratorio en instrumentos destinados a ser aplicados en el campo (Presser, Couper, Lessler, Martin, Martin, Rothgeb & Singer, 2004).

Con respecto a la primera interrogante, los datos muestran en efecto que las personas difieren mucho en su capacidad de expresar verbalmente lo que pasa por sus mentes a la hora de contestar un cuestionario. Un ejemplo de ello proviene del estudio de Castellano-Ackermann & Blair (2004), en el cual se evaluó la cantidad de problemas de comprensión que se pueden detectar en un cuestionario mediante EC, en 90 personas de diferentes niveles educativos. Los autores encontraron que las personas con mayores niveles educativos fueron las que detectaron más problemas semánticos en comparación con las personas de menores niveles educativos. Al parecer, quienes poseían mayores niveles educativos tenían una mejor capacidad de expresar los problemas de comprensión en los reactivos y, por tanto, los autores concluyen que el uso más eficiente de la EC se puede hacer precisamente en personas de altos niveles educativos. Sin embargo, como ya hemos discutido, la selección de las personas para la EC depende de los objetivos del estudio principal, por lo que debe ser guiada por la teoría y las experiencias previas de investigación.

En lo que respecta al segundo cuestionamiento (¿difiere la EC sustancialmente de los métodos tradicionales ¿), los resultados son mixtos, algunas veces contradictorios y en todo caso imposibles de comparar debido a que se utilizan distintas técnicas de mejoramiento de instrumentos y distintos criterios a la hora de identificar los problemas. Por ejemplo, Huges (2003) encontró que la EC y la técnica de *debriefing* fueron más efectivas a la hora de identificar problemas de comprensión de las preguntas de una encuesta sobre uso de computadoras en comparación con la observación del comportamiento de los entrevistados a la hora de llenar la encuesta (conocida como codificación conductual). Por otro lado, Presser & Blair (1994) encontraron que los paneles de expertos fueron los más productivos para detectar problemas semánticos en encuestas sobre salud y nutrición en comparación con técnicas tradicionales de revisión de cuestionarios, con la codificación conductual y con la misma EC.

En lo que respecta a la pregunta sobre la efectividad de esta técnica, a la fecha identificamos un solo programa de investigación dedicado a estudiar experimentalmente el impacto de las técnicas cualitativas sobre los cuestionarios utilizados en el campo. Se trata del trabajo de Barbara Forsyth, Jennifer Rothgeb y Gordon Willis (Rothgeb, Willis, & Forsyth, 2004, 2007).

En estos estudios, los autores evaluaron un cuestionario que contenía preguntas, sobre tópicos como el medio ambiente o el transporte público, mediante diferentes técnicas cualitativas aplicadas en el laboratorio, entre las que se incluía la EC. Una vez realizada la revisión, los autores desarrollaron una versión mejorada del cuestionario y una versión sin mejorar. Posteriormente, realizaron un experimento de campo en donde aplicaron las dos versiones del cuestionario de manera aleatoria en una encuesta telefónica,

y registraron los problemas que se presentaron durante la encuesta utilizando diversos criterios de calidad de las preguntas, como la cantidad de “no respuestas”, las dudas de los encuestados y las apreciaciones de los mismos encuestadores.

Los resultados indicaron que aquellos reactivos identificados como problemáticos durante la revisión cualitativa también presentaron problemas durante la entrevista telefónica, mostrando que los problemas detectados en el laboratorio pueden predecir los potenciales problemas que se pueden presentar en el campo. Sin embargo, sus resultados están muy lejos de ser concluyentes, porque en muchos aspectos el cuestionario mejorado no se diferenció del cuestionario sin mejorar. Por ejemplo, el porcentaje de “no respuestas” no se disminuyó significativamente utilizando la versión mejorada.

Ahora bien, no conocemos ningún estudio hasta ahora publicado que se haya dedicado a analizar el impacto de la EC sobre las características psicométricas de instrumentos de medición de actitudes. Los estudios que acabamos de describir se dedicaron a estudiar la calidad de preguntas de hechos y opiniones, pero el estudio de las escalas de actitud requiere de la aplicación de los principios psicométricos revisados en la primera sección de este cuaderno para estimar la pertinencia, confiabilidad y validez de las interpretaciones que hagamos con ellas. Se trata de una tarea difícil, puesto que las diferentes técnicas producen escalas relativamente diferentes. Afortunadamente, la Psicometría moderna nos ofrece diversas herramientas para comparar escalas de distinta composición que tratan de medir el mismo constructo.

Tomando en cuenta este vacío en la investigación, nos dimos a la tarea de diseñar un estudio explícitamente dedicado a examinar el efecto de la EC en las propiedades psicométricas de una escala de actitudes hacia la homofobia. En las siguientes líneas describimos dicho estudio.

El impacto de la EC en las características psicométricas de una escala de actitudes

La idea general de nuestro estudio era determinar el potencial aporte de la EC en las características psicométricas de instrumentos de medición de actitudes. Para ello comparamos las características psicométricas de una escala mejorada mediante EC con las propiedades de la misma escala mejorada mediante el juicio de un panel de expertos y las características psicométricas de la misma escala, pero sin mejorar. Como nuestro objetivo específico era comparar la confiabilidad y validez de las tres versiones, realizamos un experimento de campo para obtener evidencias de confiabilidad y validez, tal y como se describe en la primera sección de este cuaderno metodológico.

Diseño general del estudio

Como ya indicamos en la sección anterior, nuestro estudio inició con el desarrollo de una escala experimental por parte de estudiantes avanzados de psicología. Ya indicamos que los y las estudiantes desarrollaron 42 reactivos tendientes a medir actitudes negativas hacia la homosexualidad y que mediante la aplicación de 10 entrevistas cognitivas se detectaron diversos problemas. A partir de las entrevistas mejoramos los reactivos defectuosos,

eliminamos reactivos que no podían ser mejorados y desarrollamos otros nuevos, con lo cual llegamos obtuvimos una versión mejorada de 37 reactivos.

Simultáneamente realizamos un panel de expertos. Este se llevó a cabo en una sola sesión de aproximadamente dos horas y media de duración, en donde los expertos nos ayudaron a mejorar la escala original. Participaron como expertos un psicólogo, una socióloga, una filóloga y un filósofo. Todos ellos contaban con experiencia, ya sea en el constructo de la homofobia, o bien en el diseño de instrumentos de medición social. Un mes antes de la sesión, a los jueces se les envió la escala original con la única consigna de revisarla para que en la sesión de trabajo pudiéramos mejorarla.

Los jueces nos ofrecieron muchos consejos para mejorar la escala. Por ejemplo, nos recomendaron eliminar aquellos reactivos que medían acciones o conocimiento en lugar de actitud (por ejemplo, "¿Conoce usted personas homosexuales?"). También nos hicieron ver la necesidad de redactar todos los reactivos en forma de afirmaciones en primera persona y evitar preguntas de tipo "¿Considera a un hombre o mujer homosexual como una persona que es parte de una "moda" de la nueva generación?". Según los jueces, este tipo de reactivos se desajustan a las opciones de respuesta típicas de una escala Likert y son particularmente ambiguos.

Con respecto al tema de la ambigüedad, los jueces nos indicaron que debíamos evitar aquellos reactivos en donde se incluían dos o más objetos actitudinales o dos o más valencias actitudinales, es decir, dos o más adjetivos y dos o más sustantivos. Adicionalmente, nos hicieron observaciones sobre los reactivos que eran muy extensos, muy complicados o de redacción compleja. Finalmente, nos recomendaron generar más reactivos que expresaran

animadversión, disgusto, desagrado y enojo frente a la homosexualidad y a las personas homosexuales. Sobre la base de las recomendaciones de los jueces, construimos la tercera versión de la escala, la cual consta de 38 reactivos.

En general las recomendaciones de los jueces coincidieron en mucho con las recomendaciones que nos hicieron los participantes de las EC. Sin embargo, los expertos fueron particularmente enfáticos en la necesidad de incluir otros reactivos que captaran mejor el constructo, es decir, desarrollar reactivos que expresaran más fobia. En el ANEXO 2 presentamos las tres versiones de la escala, en donde se pueden apreciar los cambios realizados a los reactivos.

Una vez construidas las tres versiones, procedimos a realizar el experimento de campo. En este experimento participaron estudiantes universitarios, quienes fueron asignados aleatoriamente a una de las tres versiones de la escala, de tal manera que un grupo completó la escala sin mejorar, otro completó la escala mejorada mediante EC y un tercer grupo, la escala mejorada mediante jueces expertos.

Participantes

Trabajamos con los 310 estudiantes universitarios heterosexuales de la muestra original que contaba con una edad promedio de 21 años (DE = 4,33 años). El 98% de ellos eran costarricenses, el 66% eran mujeres y el 95% no tenía pareja a la hora de llenar la encuesta. Los y las estudiantes pertenecían a las 13 facultades que componen la Universidad de Costa Rica y todos estudiaban en el Campus Rodrigo Facio. Cien de los estudiantes completaron la escala sin mejorar, 105 la escala mejorada mediante EC y 105 la escala mejorada mediante entrevista cognitiva.

Instrumentos

Como nuestro objetivo involucraba recopilar diversas evidencias de validez, debimos incluir otras escalas y medidas para tal fin. Para estimar la validez convergente incluimos una escala de reconocida validez que midiera el mismo constructo de nuestras escalas (homofobia). Para estimar la validez discriminante incluimos dos escalas que medían constructos distintos al de la homofobia. Para estimar la validez de criterio incluimos medidas de criterios externos a la ejecución de las escalas. Así, cada estudiante completó una de las tres versiones de nuestra escala de homofobia, todas las escalas y medidas de validación y el módulo de variables sociodemográficas.

Como criterio de validez convergente elegimos la Escala de Homofobia Moderna (EHM) de Raja y Stockes adaptada al contexto latinoamericano por León (2003). Esta escala consta de 41 reactivos tendientes a medir actitudes negativas hacia las personas homosexuales. Algunos ejemplos de los reactivos de esta escala son: “No entablo relaciones amistosas con un gay porque temo al contagio del sida”, “Los psicólogos y los psiquiatras deberían esforzarse en encontrar una cura para la homosexualidad masculina” o “Los matrimonios entre gays deberían ser legalizados” (ítem inverso).

Para expresar su rechazo o apoyo a cada reactivo, los y las estudiantes contaban con una escala Likert de 5 puntos de 1 (totalmente en desacuerdo) a 5 (totalmente de acuerdo). Una vez recodificados los reactivos inversos, procedimos a calcular el promedio de las respuestas a los 41 reactivos de tal manera que altas puntuaciones indicaran altos niveles de rechazo u homofobia.

Como criterios de validez discriminante elegimos la escala de Deseabilidad social de Crowne & Marlowe (1960) y el PANAS (Positive and Negative Affect Schedule) de Watson & Clark (1994).

La versión que utilizamos de la escala de Deseabilidad Social incluye reactivos que describen conductas altamente deseables, pero que (siendo honestos) casi ninguno de nosotros es capaz de hacer ("Siempre acepto cuando me equivoco. Siempre acepto mis errores"). También incluye reactivos que describen conductas condenadas socialmente pero que (siendo honestos también) todos hacemos ("Han habido ocasiones en las que siento envidia por la buena suerte que otros tienen").

Este tipo de reactivos permiten medir la tendencia de las personas a responder de manera esperada socialmente. Para nosotros era importante incluir esta medida debido a que la homofobia es una reacción altamente condenada en contextos académicos muy "políticamente correctos".

Cada una de las situaciones descritas en los reactivos era respondida en un formato de "falso y verdadero". Asignamos un punto por cada conducta socialmente deseable contestada como verdadera y un punto por cada conducta socialmente condenada contestada como falsa, de tal manera que al sumar todos los puntos, un mayor puntaje indicara una mayor tendencia a convenir con las normas sociales, reportando con mayor facilidad conductas socialmente esperadas que conductas ilícitas y socialmente sancionadas.

El PANAS, por su parte, es una medición del estado de ánimo de las personas antes de llenar el cuestionario y es particularmente idónea para estimar en qué medida las respuestas de las personas están afectadas por su estado de ánimo y no por el objeto actitudinal en cuestión. El PANAS incluye 20

estados de ánimo, 10 positivos (“interesado/a”, “alerta”, “inspirado/a”, “atento/a”) y 10 negativos (“nervioso/a”, “angustiado/a”, “confundido/a”, “irritable”). Los y las estudiantes reportaban en una escala de 1 (no del todo) a 5 (mucho) en qué medida se habían sentido de estas maneras en las últimas dos semanas. Calculamos dos tipos de estados de ánimos: el positivo a partir del promedio de respuestas a todos los estados de ánimo positivos y el negativo a partir del promedio de respuestas de todos los reactivos negativos.

Como criterio externo utilizamos preguntas directas sobre la cantidad de amigos y amigas homosexuales que tienen los encuestados en la actualidad, la frecuencia y calidad del contacto que tienen con personas homosexuales (Smith-Castro, 2003) y su opinión (a favor o en contra) sobre el proyecto de ley en torno al matrimonio entre personas del mismo sexo. Incluimos estos criterios externos a la ejecución de las escalas por cuanto las teorías sobre la homofobia y los estudios previos indican que la homofobia está vinculada al poco contacto con personas homosexuales y a una tendencia a rechazar las reivindicaciones de este colectivo (Smith-Castro & Molina, en rev.).

Para asegurar la calidad de todas estas medidas realizamos un estudio de validación, el cual llevamos a cabo en el primer ciclo del 2009 en una muestra de 107 estudiantes de la Escuela de Estudios Generales de la UCR. Los resultados de este estudio evidenciaron que los instrumentos de validación poseen excelentes índices de consistencia interna y se relacionan con los otros constructos de la manera esperada por la teoría, lo cual evidencia su capacidad para servir como criterios de validación de nuestras escalas (Smith-Castro et al., 2009).

Procedimientos

Una vez obtenidos los respectivos permisos, contactamos a los estudiantes en sus aulas universitarias, les explicamos el objetivo del estudio y les pedimos contestar el cuestionario en su totalidad. Siguiendo los lineamientos del Comité de Ética de nuestra universidad, incluimos la fórmula de consentimiento informado en la portada de nuestros cuestionarios. Los y las estudiantes expresaron su acuerdo en participar del estudio completando el cuestionario en su totalidad. Quien no deseaba participar era libre de devolver el cuestionario vacío. A los y las estudiantes les tomó aproximadamente 30 minutos llenar el cuestionario.

Procesamos los datos utilizando hojas de cálculo de Excel y el paquete estadístico SPSS 17. Realizamos los análisis clásicos de consistencia interna, análisis de ítems y correlaciones bivaridas entre las escalas en estudio y las escalas de validación para estimar estadísticamente la confiabilidad y validez.

Resultados

Presentamos los resultados en dos grandes dimensiones. Primero analizaremos los resultados de los análisis de consistencia interna y las propiedades de los reactivos y posteriormente presentamos las correlaciones entre nuestras escalas y los criterios de validación.

Evidencias de consistencia interna

Los resultados del análisis de consistencia interna se presentan en el Cuadro 18. Para estimar la consistencia interna de las escalas utilizamos el índice Alfa de Cronbach. Este índice oscila entre 0 y 1, en donde 1 representa

una consistencia interna perfecta. En la investigación social, índices superiores a .80 son indicadores de consistencias internas adecuadas.

Cómo se puede observar en el Cuadro 18, todas las escalas presentaron índices de consistencia interna excelentes, pero las escalas mejoradas (y en particular mediante la EC) presentaron índices de consistencia interna más robustos que la escala sin mejorar.

Cuadro 18. Resultados del análisis de consistencia interna de las escalas.

	Sin mejorar	EC	Jueces
Alfas de Cronbach	.90	.94	.92
Número de reactivos	42	37	38
Promedio de las correlaciones ítem-total	.40	.52	.48
% de ítems que presentan correlaciones ítem-total inferiores a .30	31% (n = 13)	14% (n = 5)	21% (n = 8)

Existe la posibilidad de estimar si tales diferencias son esperables por puro azar o bien son diferencias estadísticamente significativas. Se trata de la prueba Hakstian-Whalen (1976) para estimar las diferencias entre dos o más índices de consistencia interna Alfa de Cronbach. Esta prueba toma en cuenta la cantidad de reactivos que componen cada escala y la cantidad de sujetos que la contestaron para producir un estadístico que refleja la diferencia entre los índices.

Aplicando esta prueba nos dimos cuenta de que la escala mejorada mediante EC presente un índice de consistencia Alfa de Cronbach significativamente mayor que el que presenta la escala sin mejorar (χ^2 = 6,07, gl = 1, *p* = .014), pero que las dos escalas mejoradas (ya sea mediante EC o mediante jueces expertos) no difieren significativamente entre sí (χ^2 = 1,75, gl = 1, *p* = .18). Además, la consistencia interna de la escala mejorada mediante jueces expertos tampoco difiere significativamente de la consistencia interna de la escala sin mejorar (χ^2 = 1.05, gl = 1, *p* = .30).

Además de analizar la consistencia interna de las escalas, examinamos las características de los reactivos que las componen mediante el índice de discriminación. Como ya indicamos, este índice informa sobre la capacidad de los reactivos de diferenciar entre los niveles altos y bajos del constructo que pretenden medir, y se define como la correlación simple entre las puntuaciones de los sujetos en el reactivo y sus puntajes en la escala como un todo (Muñiz, 2002). Por lo general, cuando un reactivo presente un índice de discriminación menor a .30 se considera que no cumple bien con su función de discriminar, ya que se encuentra pobremente asociado al total de la escala.

Examinando cada una de las escalas, nos dimos cuenta de que la cantidad de reactivos con índices de discriminación menores a .30 era mayor para la escala sin mejorar que para las escalas mejoradas. Específicamente encontramos que en la escala sin mejorar, un 31% de los reactivos (13 de 42) no poseían propiedades de discriminación adecuadas. Se trata de reactivos que hubiéramos tenido que eliminar para asegurarnos una mayor confiabilidad. En cambio, el porcentaje de reactivos con poca capacidad de discriminación

disminuyó al 21% (8 de 38) en la escala mejorada mediante jueces expertos y a un 14% (5 de 37) en el caso de la escala mejorada mediante EC.

En suma estos primeros análisis nos indican que las escalas mejoradas, particularmente mediante EC, poseen una mayor consistencia interna que la escala sin mejorar y sus reactivos poseen mejores índices de discriminación.

Evidencias de validez

Recordemos que en este estudio tratamos de recopilar evidencias de validez convergente, discriminante y de criterio, incluyendo en el cuestionario varias escalas de validación.

Las evidencias sobre la validez convergente de "nuestras" escalas estarían proporcionadas si las puntuaciones de los sujetos que se derivan de ellas se correlacionan positivamente con la escala de homofobia de consabida validez (pues miden lo mismo). Por otro lado, obtendríamos importantes evidencias de validez discriminante, si las puntuaciones no se correlacionan con la escala de Deseabilidad Social y el PANAS, lo que indicaría que no estamos confundiendo constructos y que las respuestas no se ven afectas por el particular estado de ánimo de la persona a la hora de contestar el cuestionario o por su tendencia natural a convenir con las normas.

Adicionalmente, obtendríamos importantes evidencias de la validez de criterio si observamos que altos puntajes en homofobia se correlacionan negativamente con la cantidad de amigos y amigas homosexuales y la cantidad y calidad de contacto con personas homosexuales. Esperaríamos, además, que altos puntajes en homofobia estén correlacionados con una negativa a apoyar el matrimonio o la unión civil entre personas homosexuales.

Para estimar la asociación entre las variables utilizamos la correlación simple o correlación de Pearson. El índice mide el grado de relación de dos variables cuantitativas (métricas). El coeficiente presenta valores entre –1 y +1. Cuando el índice se aproxima a cero, indica que no existe correlación lineal entre las variables, si se aproxima a +1 indica la existencia de una correlación positiva (directamente proporcional) entre las variables y si se aproxima a -1 indica que existe una correlación negativa (inversamente proporcional) entre las variables. Estos resultados se presentan en el Cuadro 19.

Cuadro 19. Correlaciones de las escalas con los criterios de validez.

	Sin mejorar	Entrevista Cognitiva	Jueces Expertos
Deseabilidad social (Crowne & Marlowe, 1960)	.01	.02	-.12
Mood negativo (Watson & Clark, 1994)	.13	-.10	-.01
Mood positivo (Watson & Clark, 1994)	-.03	.05	-.06
Homofobia (León, 2003)	.87 **	.91 **	.91 **
Contacto (Smith-Castro, 2003)	-.53 **	-.51**	-.56 **
Calidad de contacto	-.62 **	-.47 **	-.63 **
Número de amigos y amigas gay	-.36 **	-.31 **	-.39 **
Unión civil gay (0 = no, 1 = si)	-.62 **	-.62 **	-.57 **
Matrimonio gay (0 = no, 1 = si)	-.60 **	-.62 **	-.53 **

** $p < .001$

El Cuadro 19 nos presenta importantes evidencias sobre la validez de las inferencias que podemos hacer con nuestras escalas. Por ejemplo, ninguna correlaciona significativamente con la tendencia a convenir con las normas sociales y ninguna se ve afectada significativamente por el estado de ánimo de las personas antes de contestarlas. Por otra parte, las escalas correlacionan

positivamente con otra escala de reconocida validez de homofobia, lo cual indica que estamos midiendo el mismo constructo. Además, todas las escalas se relacionan con los criterios externos tal y como la teoría lo predice, presentando correlaciones negativas con el contacto y las actitudes hacia las iniciativas de la unión civil y el matrimonio gay: altos niveles de homofobia se asociaron a pocas oportunidades de contacto con personas homosexuales y a una resistencia a apoyar el matrimonio y la unión civil entre personas del mismo sexo.

Las correlaciones entre nuestras escalas y los criterios de validación son tan consistentes y similares que no es posible dictaminar cuál de ellas se correlaciona "mejor" con los criterios. De hecho, existen pruebas de significancia para estimar si dos correlaciones son distintas entre sí, más allá de lo esperable por puro azar (Fisher, 1921), y en todos nuestros casos no encontramos ninguna diferencia estadísticamente significativa entre las correlaciones (todas las $p > .05$).

En síntesis, los análisis que hemos realizado arrojan resultados bastante halagadores para la técnica de EC. La escala mejorada mediante EC se comporta tan eficazmente como una escala mejorada mediante un panel de expertos, pero a diferencia del panel expertos, la EC permite obtener información sobre cómo los sujetos (la población meta) están entendiendo los reactivos.

Si bien todas las escalas correlacionan con los criterios de validación tal y como se espera de acuerdo con la teoría, la escala mejorada mediante EC posee propiedades de consistencia interna significativamente superiores. La escala sin mejorar, por su parte, requiere que los sujetos contesten más

reactivos (y por tanto inviertan más tiempo y más recursos cognitivos) para obtener la misma información que con las escalas mejoradas. Y aún poseyendo más reactivos, su consistencia interna es menor.

Literatura recomendada

Esta sección puede considerarse como un pequeño ejemplo del proceso de validación de una escala para medir actitudes, pues tratamos de seguir los pasos más básicos para recopilar evidencias sobre la consistencia interna y la validez de un instrumento, tal y como recomiendan los estándares modernos que describimos en la primera sección. Así, la literatura que recomendamos en la primera sección resulta particularmente idónea para profundizar en algunos aspectos técnicos de esta sección.

Nuestro estudio presenta las formas más simples de estimar evidencias de confiabilidad y validez. Existen procedimientos más complejos y más robustos para ello. El texto de Tornimbeni, Pérez y Olaz (2008) presenta más y mejores ejemplos sobre los procedimientos para estimar la confiabilidad y la validez de las mediciones que los que ofrecemos aquí.

Para quienes deseen mayor información sobre cómo calcular pruebas de hipótesis sobre las diferencias entre coeficientes Alfas de Cronbach, recomendamos el texto de Hasktian y Whalen (1976). El profesor Hoi K. Suen, de la Facultad de Educación de la Universidad Estatal de Pensilvania, ofrece gratuitamente una hoja de Excel para hacer los respectivos cálculos. Esta hoja puede ser solicitada a la siguiente dirección de correo electrónico: HoiSuen@psu.edu

Para conocer más a fondo los distintos procedimientos de estimación de las diferencias entre dos o más coeficientes de correlación, así como para

acceder a muchas otras estrategias de análisis estadístico, recomendamos visitar la página del profesor Karl Wunch de la Universidad de Carolina del Este: http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm

UNAS CUANTAS RECOMENDACIONES FINALES

La EC fue diseñada para detectar problemas en los cuatro grandes momentos que se suceden a la hora de contestar un cuestionario de papel y lápiz, a saber: comprensión de los reactivos, recuperación de la información solicitada, proceso de estimación de las respuestas y ejecución de las respuestas.

Se trata de un procedimiento de naturaleza cualitativa que informa al investigador o investigadora sobre cómo las personas procesan la información solicitada por los reactivos, pero también produce importantes evidencias de que el instrumento está midiendo lo que pretende medir y de que lo hace de manera consistente.

Nuestra experiencia indica que se trata de una técnica flexible, que puede ser adaptada a las necesidades de cada cuestionario, estudio o equipo de investigación. Hemos visto que la EC puede ser aplicada con éxito en poblaciones tan diferentes como lo son niños y niñas menores de 12 años, personas refugiadas y estudiantes universitarios.

Esta flexibilidad aumenta si las pruebas de reporte verbal son vistas como un catálogo de posibilidades para el sondeo de los mecanismos mentales implicados a la hora de contestar nuestros cuestionarios y de los problemas que pueden aparecer durante el proceso de pregunta-respuesta.

Su aplicación es relativamente sencilla y poco costosa, porque requiere un equipo mínimo y una muestra modesta de participantes. En este sentido, la selección de los participantes en esta fase del proceso de desarrollo de

instrumentos resulta crucial. Se trata de una selección guiada teóricamente y por los objetivos de investigación ulteriores.

No cabe duda que mejorar cualitativamente un instrumento antes de pasar a las siguientes etapas del proceso de validación tiene un impacto positivo en el producto final. Como hemos visto, la aplicación de la EC permite mejorar sustancialmente las características psicométricas de los instrumentos de papel y lápiz.

No es de extrañar entonces que la EC haya sido utilizada en los últimos 30 años como herramienta para evaluar y mejorar los cuestionarios antes de probarlos en los estudios piloto y llevarlos al estudio principal. De hecho, las principales casas encuestadoras en Estados Unidos y Europa poseen laboratorios permanentes de evaluación cognitiva de sus instrumentos (Willis, 2005).

Así mismo, la Asociación Americana de Investigación en Educación (AERA, por sus siglas en inglés), la Asociación Americana de Psicología (APA) y el Consejo Nacional de Medición en Educación (NCME) de Estados Unidos recomiendan el uso de técnicas como la EC para recopilar evidencias de validez del proceso de respuesta como parte fundamental del desarrollo, validación y aplicación de pruebas psicométricas y educativas (AERA, APA & NCME, 1999). No podemos más que unirnos a estas instituciones en la recomendación de su uso.

Hemos visto que su uso más intensivo se ubica en las encuestas sobre hechos y conductas (como por ejemplo, uso del condón, consumo de tabaco, frecuencia de visitas al médico, etc.). No obstante, puede ser fácilmente llevada al análisis de reactivos tendientes a medir valores, actitudes y personalidad,

como lo hemos tratado de demostrar en este cuaderno metodológico y como ha sido incorporada para desarrollar instrumentos de habilidades y aptitudes (Embretson & Gorin, 2001).

Ahora bien, la EC en sí misma no sustituye a un estudio psicométrico de validación de reactivos, sino que forma parte de los procedimientos modernos de estimar la validez y utilidad de los instrumentos, y en ese sentido la recomendamos como un excelente complemento a los métodos tradicionales de validación y adaptación de medidas y recomendamos su uso en las etapas iniciales del proceso de construcción de instrumentos, antes de llevarlas al estudio piloto de carácter psicométrico y por supuesto antes del estudio principal.

En general la EC posee características sumamente deseables para la investigación social. Nuestros datos indican que es tan eficiente como un buen panel de expertos, pero tiene la ventaja de que nos informa sobre lo que sucede en las mentes de los potenciales encuestados, que al fin y al cabo representan nuestras unidades de estudio centrales. De esta manera podríamos decir que ambas estrategias son complementarias. Los jueces expertos nos ayudaron a afinar los reactivos para que captaran mejor el constructo, mientras que la EC nos permitió estar seguros de que el proceso mental involucrado a la hora de contestar nuestros reactivos era el adecuado (para nuestros fines).

Finalmente, una de las características más relevantes de esta técnica es que nos obliga a pensar críticamente sobre nuestros instrumentos. Nos recuerda que todas nuestras medidas son imperfectas y que no podemos dar

por sentadas nuestras expectativas (o cualquier otra teoría) sin someterlas primero a un proceso de investigación empírica, rigurosa y transparente.

Invitamos entonces a incorporar la EC dentro del repertorio de herramientas de investigación social para que la toma de decisiones dentro y fuera del ámbito académico esté fundamentada en mediciones fiables y efectivas.

Sobre los autores

Vanessa Smith-Castro

Doctora en Psicología por la Universidad Philipps de Marburgo, Alemania. Actualmente es profesora asociada del Instituto de Investigaciones Psicológicas y la Escuela de Psicología de la Universidad de Costa Rica. Dentro de sus áreas de interés se encuentran la psicología social de las relaciones intergrupales y los métodos de investigación cuantitativa.

Correo electrónico: vanessa.smith@ucr.ac.cr

Mauricio Molina Delgado

Doctor en Psicología por la Universidad de Salónica, Grecia. Catedrático de la Universidad de Costa Rica, director de la Maestría en Ciencias Cognoscitivas, profesor en la Escuela de Psicología e Investigador del Instituto de Investigaciones Psicológicas y el Programa de Investigación en Neurociencias, en esta misma Universidad. Dentro de sus intereses de investigación se encuentran las ciencias cognoscitivas, la simulación de comportamiento animal, el procesamiento de lenguaje no literal y la metacognición.

Correo electrónico: orescu@yahoo.com

REFERENCIAS

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.

Aguiar, T. (2009). *Adaptación de la escala de neuroticismo INK para estudiantes de cuarto, quinto y sexto grado.* Tesis para optar por el grado de Licenciatura en Psicología. San José, C.R.: Universidad de Costa Rica.

Anastasi, A. (1982*). Psychological testing* (5th ed.). New York, NY: Macmillan.

Bollen, K.A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology, 53,* 605-34.

Bradburn, N. M., Rips, L.J. & Shevell, S. K. (1987). Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. *Science, 236 (4798),* 157-61.

Castellano-Ackermann, A. & Blair, J. (May, 2006). *Efficient Respondent Selection for Cognitive Interviewing.* Paper presented at the 61st Annual Conference of the American Association for Public Opinion Research. Montreal, Quebec.

Collins, D. (2003). Pretesting survey instruments: an overview of cognitive methods. *Quality of Life Resereach, 12,* 229-238.

Conway, M.A. & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self memory system. *Psychological Review, 107,* 261-288.

Cortada de Kohan, N. (2003). Posibilidad de integración de las teorías cognitivas y la psicometría moderna. *Revista Argentina de Neuropsicología, 1,* 8 –23.

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Crowne, D. P. & Marlowe, D. (1960). A New Scale of Social Desirability Independent of Psychopathology*, Journal of Consulting Psychology, 24,* 349-354.

Embretson, S. & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38(4),* 343-368.

Erickson, K.A. & Simon, H.A. (1993). *Protocol Analysis. Verbal reports as data (Revised edition).* Cambridge, MA: MIT Press.

Fisher, R. A. (1921). On the ``probable error'' of a coefficient of correlation deduced from a small sample. *Metron 1,* 3-32.

Flik, U., Kardorff, E., Keupp, H., Rosentiel, L., & Wolff, S. (1995). *Handbuch qualitative Sozialforschung: Grundlangen, Konzepte, Methoden und Anwendungen*. Weinheim: Beltz Psychologie Verlags Union.

Forsyth, B., Rothgeb, J., and Willis, G. (2004). Does Pretesting Make a Difference? An Experimental Test. In S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds.), *Methods for Testing and Evaluating Survey Questionnaires,* pp. 525-546. New York: Wiley.

Galton, F. (1879). Psychometric experiments. *Brain*, *2*, 149 – 162.

Hakstian, A. R. & Whalen, T. E. (1976). A k-sample significance test for independent Alpha coefficients. *Psychometrika , 41*, 219-231.

Huges, K.A. (November, 2003). *Comparing Pretesting Methods: Cognitive Interviews, Respondent Debriefing, and Behavior Coding*. Paper presented at the Annual Meeting of the Federal Committee on Statistical Methodology. Arlington, Virginia.

Jobe, J.B. & Mingay, D.J. (1991). Cognition and Survey Measurement: History and Overview. *Applied Cognitive Psychology, 5,* 175-192.

Jobe, J.B. (2003). Cognitive psychology and self-reports: Models and Methods. *Quality of Life Research, 12,* 219-227.

Kaplan, R. M. & Sacuzzo, D. P. (2006). *Pruebas psicológicas. Principios, aplicaciones y temas* (6° Ed.). México: International Thomson Editores.

Klein, S.B. German, T.P., Cosmides, L. & Gabriel, R. (2004). A theory of autobiographical memory: necessary components and disorders resulting from their loss. *Social Cognition, 22,* 460-490.

Lessler, J., Tourangeau, R. & Salter, W. (1989). *Questionnaire design in the cognitive research laboratory: results of an experimental prototype*. National Center for Health Statistics. *Vital Health Stat*, *6*(1), Washington, D.C.: Government Printing Office.

Lewins, A. & Silver, Ch. (2007). *Using Software in Qualitative Research: A Step-by-Step Guid*e. California: Sage Publications.

Mack, N., Woodsong, C., McQueen, K., Guest, G. & Namey, E. (2005). *Qualitative Research Methods: A Data Collector's Field Guide*. North Carolina: Familiy Health International.

Martínez, R. (1996). *Psicometría: Teoría de los Tests Psicológicos y Educativos.* España: Editorial Síntesis S.A.

Mayring, P. (1993). *Einfuhrung in die qualitative Sozialforschung.* Weinheim: Beltz Psychologie Verlags Union.

McIntire, S.A. & Miller, L.A. (2007). *Foundations of Psychological Testing. A Practical Approach* (Second Edition). California, United States of America: Sage Publications, Inc.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5-11.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Montero, E. (2001). La teoría de respuesta a los ítemes: una moderna alternativa para el análisis psicométrico de instrumentos de medición. *Revista de Matemática: Teoría y Aplicaciones, 7,* 217-228.

Montero, E. (2008). Escalas o Índices para la medición de constructos: El dilema del analista de datos. *Avances en Medición, 6*, 15–24.

Moreira, T. (2008). Funcionamiento diferencial del ítem en matemática. Un aporte teórico metodológico. *Actualidades en Psicología, 22,* 91-113.

Norenzayan, A., & Schwarz, N. (1999). Telling what they want to know: Participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology*, 29, 1011-1020

Neuendorf, K.A. (2002). *The content analysis guidebook.* Thousand Oaks, Cal.: Sage.

Nunnally, J. (1991). *Teoría Psicométrica*. México: Trillas

Osterlind, S. J. (2006). M*odern measurement: Theory, principles and application of mental appraisal*. Upper Saddle River, NJ: Prentice-Hall.

Padilla, J.L., Gómez, J., Hidalgo, M.D. & Muñiz, J. (2006). La evaluación de las consecuencias del uso de los tests en la teoría de la validez. *Psicothema*, *18*(2), 307-312.

Phelps, E.A. (2006). Emotion and cognition: Insights from studies of the human amygdala. *Annual Review of Psychology, 24*, 27-53.

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results? *Sociological Methodology*, *24*, 73-104.

Presser, S., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Rothgeb, J.M. & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68,* 109-130

Rothgeb, J., Willis, G., & Forsyth, B. (2007). Questionnaire pretesting methods: Do different techniques and different organizations produce similar results. *Bulletin of Sociological Methodology, 96,* 5-31.

Ryle, G. (1949). *The concept of mind*. New York: Barnes & Noble.

Schawarz, N. & Oyserman, D. (2001). Asking questions about behavior: cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22,* 127-160.

Smith, A.F., Jobe, J.B. & Mingay, D.J. (1991). Questions induced cognitive biases in reports of dietary intake in college men and women. *Health Psychology, 10(4)*, 244-257.

Smith-Castro, V. (2003). *Acculturation and psychological adaptation.* Westport, CT: Greenwood Press.

Smith-Castro, V. & Molina, M. (en rev.). Matrimonio Gay en Costa Rica: ¿Autoritarismo, homofobia o desconocimiento? Manuscrito en revisión. *Revista Interamericana de Psicología.*

Smith-Castro, V., Molina, M. & Castelain, R. (2009). *Nuevos métodos y tecnologías lingüístico-cognitivas para el diseño, evaluación y mejoramiento de cuestionarios en la investigación social*. Informe de Investigación. San José, Costa Rica: Instituto de Investigaciones Psicológicas.

Thurstone, L.L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33,* 529-554.

Tirapu-Ustárroz, J., Muñoz-Céspedes, J.M. (2005). Memoria y funciones ejecutivas. *Revista de Neurología, 41* (8), 475-484.

Tornimbeni, S., Pérez, E. & Olaz, F. (2008). *Introducción a la Psicometría*. Buenos Aires: Paidós.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In Jabine T.B., Straf, M.L., Tanur, J.M. & Tourangeau, R (eds). *Cognitive Aspects of*

Survey Methodology: Building a Bridge between Disciplines (pp. 73-101). Washington, DC: National Academy Press.

Tourangeau, R., Rips L.J. & Rasinski, K. (2004). *The Psychology of Survey Report.* Cambridge: University Press.

Tourangeau R. & Smith, T.M. (1996). Asking sensitive questions: the impact of mode, question format, and question context. *Public Opinion Quartely, 60,* 275-304

Tulving, E. (2002). Episodic Memory: From Mind to Brain. *Annual Review of* Psychology, 53, 1-25.

Watson, D. & Clark, L.A. (1994). *The PANAS-X. Manual for the positive and negative affect schedule.* Iowa: University of Iowa.

Wengraf, T. (2006). *Qualitative research interviewing.* Thousand Oaks, Cal.: Sage.

Willis, G. (2005). *Cognitive Interviewing: A tool for improving questionnaire design.* London: SAGE.

Zúñiga M.E. & Montero, E. (2007). Teoría G: un futuro paradigma para el análisis de pruebas psicométricas. *Actualidades en Psicología, 21,* 117-144.

ANEXOS

ESCALA EXPERIMENTAL

POR FAVOR MARQUE CON UNA X SI USTED ESTÁ EN DESACUERDO O DE ACUERDO CON CADA FRASE. UTILIZANDO LA ESCALA DE 1 AL 5, EN LA QUE 1 SIGNIFICA TOTALMENTE EN DESACUERDO Y 5 TOTALMENTE DE ACUERDO. RECUERDE: CUANTO MAYOR SEA EL PUNTAJE QUE USTED ASIGNE A CADA FRASE, MAYOR ES SU ACUERDO CON LA MISMA.

ABREVIATURAS

TD: Totalmente en desacuerdo
DE: En desacuerdo
I: Indiferente
DA: De acuerdo
TA: Totalmente de acuerdo

	TD	DE	I	DA	TA
1. Asisto a lugares en donde personas homosexuales generalmente se reúnen.	1	2	3	4	5
2. Si fuera jefe(a) de una empresa le daría trabajo a una persona que es abiertamente homosexual	1	2	3	4	5
3. Me gusta ver películas de corte gay o lésbico.	1	2	3	4	5
4. ¿Considera a un hombre o mujer homosexual como una persona que es parte de una "moda" de la nueva generación?	1	2	3	4	5
5. Apoyo el matrimonio entre personas del mismo sexo.	1	2	3	4	5
6. Cuando conozco a una pareja homosexual generalmente pienso quien hace el rol de hombre o de mujer en la relación.	1	2	3	4	5
7. Podría aceptar que exista un miembro homosexual dentro de mi familia.	1	2	3	4	5
8. Continuaría una relación íntima o amorosa con una persona aunque esta me confiese haber tenido novios(as) de su mismo sexo.	1	2	3	4	5
9. Tengo amigos(as) homosexuales.	1	2	3	4	5
10. Me interesan los temas relacionados a la homosexualidad.	1	2	3	4	5
11. Considero que los homosexuales hombres y mujeres merecen en todo sentido igual de respeto que las personas heterosexuales.	1	2	3	4	5
12. Me sentiría incómodo(a) si tuviera que trabajar de cerca con una persona homosexual.	1	2	3	4	5
13. Me sentiría desilusionado(a) si tuviera un hijo(a) homosexual	1	2	3	4	5
14. Considero que las personas nacen con una predisposición a la homosexualidad.	1	2	3	4	5
15. Considero válido que dos personas del mismo sexo que tengan una relación de pareja puedan adoptar hijos.	1	2	3	4	5
16. La homosexualidad es producto de un problema genético.	1	2	3	4	5
17. Me molestaría si una persona de mi mismo sexo me invitara a salir.	1	2	3	4	5
18. Considero que la homosexualidad es una enfermedad mental o un problema, que tiene que ser tratado.	1	2	3	4	5

19. Considero que las personas homosexuales sufren por el hecho de ser homosexuales.	1	2	3	4	5
20. Si tuviera el poder de cambiar las cosas, erradicaría la homosexualidad.	1	2	3	4	5
21. ¿Considera a un hombre o mujer homosexual como una persona normal?	1	2	3	4	5
22. Considero que las personas se hacen homosexuales durante el transcurso de su vida.	1	2	3	4	5
23. Una persona que ha sufrido algún tipo de abuso sexual durante su infancia tiene altas probabilidades de convertirse en homosexual.	1	2	3	4	5
24. ¿Considera a un hombre o mujer homosexual como una persona inferior Enferma / cree que se puede curar?	1	2	3	4	5
25. Ante un hecho contra los homosexuales actuó con silencio	1	2	3	4	5
26. Aceptaría un trabajo si sabe que su jefe inmediato es homosexual.	1	2	3	4	5
27. ¿Conoce familiares o amigos homosexuales?	1	2	3	4	5
28. Usted aprobaría la unión legal en personas del mismo sexo	1	2	3	4	5
29. Usted aprobaría que un primo, sobrino o hijo recibiera clases en la escuela de un profesor con orientación sexual homosexual	1	2	3	4	5
30. La homosexualidad es una moda	1	2	3	4	5
31. Visitaría un bar gay en compañía de un amigo o amiga con orientación homosexual.	1	2	3	4	5
32. Aprobaría que las personas con orientación homosexual estuviesen en su grupo religioso.	1	2	3	4	5
33. ¿Conoce organizaciones anti-homofobia?	1	2	3	4	5
34. Las parejas o personas homosexuales pueden formar un núcleo familiar	1	2	3	4	5
35. Si ve a una mujer darle un beso a otra mujer usted está en …	1	2	3	4	5
36. Ante un hecho contra los homosexuales actuó con una denuncia	1	2	3	4	5
37. ¿Existe en C.R. la cultura para denunciar los actos ofensivos o discriminatorios por motivo de la orientación sexual?	1	2	3	4	5
38. La homosexualidad es producto del entorno social	1	2	3	4	5
39. Usted cuenta chistes acerca de la homosexualidad	1	2	3	4	5
40. Si fuera gerente de una empresa le daría trabajo a una persona homosexual.	1	2	3	4	5
41. Una persona homosexual podría ser mi compañero(a) de apartamento.	1	2	3	4	5
42. Si ve a un hombre darle la mano a otro hombre está en …	1	2	3	4	5

ESCALA MEJORADA MEDIENTA EC

POR FAVOR MARQUE CON UNA X SI USTED ESTÁ EN DESACUERDO O DE ACUERDO CON CADA FRASE. UTILIZANDO LA ESCALA DE 1 AL 5, EN LA QUE 1 SIGNIFICA TOTALMENTE EN DESACUERDO Y 5 TOTALMENTE DE ACUERDO. RECUERDE: CUANTO MAYOR SEA EL PUNTAJE QUE USTED ASIGNE A CADA FRASE, MAYOR ES SU ACUERDO CON LA MISMA.

ABREVIATURAS

TD: Totalmente en desacuerdo
DE: En desacuerdo
I: Indiferente
DA: De acuerdo
TA: Totalmente de acuerdo

	TD	DE	I	DA	TA
1. Asisto a lugares en donde hombres y mujeres homosexuales generalmente se reúnen (por ejemplo, bares gay).	1	2	3	4	5
2. Considero que las personas se hacen homosexuales durante el transcurso de su vida.	1	2	3	4	5
3. Tengo amigos(as) homosexuales.	1	2	3	4	5
4. Si fuera jefe(a) de una empresa le daría trabajo a un hombre o una mujer que es abiertamente homosexual	1	2	3	4	5
5. Me gusta ver películas de temas relacionados a la homosexualidad.	1	2	3	4	5
6. Apoyo el matrimonio por la iglesia entre personas del mismo sexo.	1	2	3	4	5
7. Considero que los hombres y las mujeres homosexuales nacen con una predisposición biológica a la homosexualidad.	1	2	3	4	5
8. Cuando conozco a una pareja homosexual generalmente pienso en quién hace el rol de hombre o de mujer en la relación.	1	2	3	4	5
9. Podría aceptar que exista un hombre o una mujer homosexual dentro de mi familia.	1	2	3	4	5
10. Me interesan los temas relacionados con la homosexualidad.	1	2	3	4	5
11. Considero que los homosexuales hombres y mujeres merecen en todo sentido igual respeto que las personas heterosexuales.					
12. Me sentiría incómodo(a) si tuviera que trabajar de cerca de un hombre o mujer homosexual.	1	2	3	4	5
13. Me sentiría desilusionado(a) si tuviera un hijo(a) homosexual	1	2	3	4	5
14. Una persona que ha sufrido algún tipo de abuso sexual durante su infancia tiene altas probabilidades de convertirse en homosexual.	1	2	3	4	5
15. Me molestaría que una persona homosexual flirteara conmigo (“me echara el cuento”)	1	2	3	4	5
16. Considero que la homosexualidad es una enfermedad	1	2	3	4	5

mental que tiene que ser tratada.					
17. Si tuviera el poder de cambiar las cosas, erradicaría la homosexualidad.	1	2	3	4	5
18. Considero a un hombre o mujer homosexual como una persona normal	1	2	3	4	5
19. Considero a un hombre o mujer homosexual como una persona que es parte de una "moda" de la nueva generación	1	2	3	4	5
20. Usted aprobaría la unión legal en personas del mismo sexo	1	2	3	4	5
21. Considero válido que dos personas del mismo sexo que tengan una relación de pareja puedan adoptar hijos.	1	2	3	4	5
22. Aprobaría que un familiar cercano recibiera clases en la escuela con un profesor homosexual	1	2	3	4	5
23. Considero a un hombre o mujer homosexual como una persona enferma	1	2	3	4	5
24. Visitaría un bar gay en compañía de un amigo o amiga con orientación homosexual.	1	2	3	4	5
25. La homosexualidad es producto de un entorno social problemático	1	2	3	4	5
26. Aprobaría que los hombres y las mujeres homosexuales asistan a grupos religiosos	1	2	3	4	5
27. Considero que los hombres y las mujeres homosexuales se sienten mal por el hecho de ser homosexuales.	1	2	3	4	5
28. Las parejas homosexuales pueden iniciar su propia familia	1	2	3	4	5
29. Se debería permitir a las parejas homosexuales adoptar hijos	1	2	3	4	5
30. La homosexualidad es una moda	1	2	3	4	5
31. Aprobaría que un familiar cercano recibiera clases en la escuela con una profesora lesbiana	1	2	3	4	5
32. Considero a un hombre o mujer homosexual como una persona inferior	1	2	3	4	5
33. La homosexualidad es producto de un problema genético	1	2	3	4	5
34. Si fuera gerente de una empresa le daría trabajo a un hombre o una mujer homosexual.	1	2	3	4	5
35. De vez en cuando cuento chistes sobre la homosexualidad	1	2	3	4	5
36. Aceptaría un trabajo si sabe que su jefe o jefa inmediato/a es homosexual.	1	2	3	4	5
37. Una persona homosexual podría ser mi compañero(a) de apartamento.	1	2	3	4	5

ESCALA MEJORADA MEDIANTE JUECES EXPERTOS

POR FAVOR MARQUE CON UNA X SI USTED ESTÁ EN DESACUERDO O DE ACUERDO CON CADA FRASE. UTILIZANDO LA ESCALA DE 1 AL 5, EN LA QUE 1 SIGNIFICA TOTALMENTE EN DESACUERDO Y 5 TOTALMENTE DE ACUERDO. RECUERDE: CUANTO MAYOR SEA EL PUNTAJE QUE USTED ASIGNE A CADA FRASE, MAYOR ES SU ACUERDO CON LA MISMA.

ABREVIATURAS

TD: Totalmente en desacuerdo
DE: En desacuerdo
I: Indiferente
DA: De acuerdo
TA: Totalmente de acuerdo

	TD	DE	I	DA	TA
1. Me incomoda ver películas o programas televisivos en donde se presentan escenas de homosexualidad	1	2	3	4	5
2. Me gusta asistir a lugares donde personas homosexuales generalmente se reúnen.	1	2	3	4	5
3. Estoy dispuesto(a) a tener amigos(as) homosexuales.	1	2	3	4	5
4. Si fuera jefe(a) de una empresa le daría trabajo a una persona que asume públicamente su condición de homosexual.	1	2	3	4	5
5. Podría aceptar que exista un miembro homosexual dentro de mi familia.	1	2	3	4	5
6. Continuaría una relación íntima o amorosa con una persona aunque esta me confiese haber tenido una pareja de su mismo sexo.	1	2	3	4	5
7. Cuando conozco a una pareja homosexual generalmente pienso quien hace el rol de hombre o de mujer en la relación.	1	2	3	4	5
8. La homosexualidad es una moda.					
9. Los homosexuales hombres y mujeres merecen en todo sentido igual respeto que las personas heterosexuales.	1	2	3	4	5
10. Me sentiría incómodo(a) si tuviera que trabajar de cerca con una persona homosexual.	1	2	3	4	5
11. Las personas nacen con una predisposición a la homosexualidad.	1	2	3	4	5
12. Me sentiría desilusionado(a) si tuviera un hijo(a) homosexual	1	2	3	4	5
13. Dos personas del mismo sexo que tengan una relación de pareja pueden adoptar hijos.	1	2	3	4	5
14. Me incomodaría mucho si una persona de mi mismo sexo me invitara a salir.	1	2	3	4	5
15. Las personas se hacen homosexuales durante el transcurso de su vida.	1	2	3	4	5
16. Las personas homosexuales sufren por el hecho de ser homosexuales.	1	2	3	4	5
17. La homosexualidad es una enfermedad mental					
18. La verdad es que yo eliminaría la homosexualidad.	1	2	3	4	5

19. Un hombre o mujer homosexual es una persona normal	1	2	3	4	5
20. De vez en cuando siento un poco de desprecio por las personas homosexuales	1	2	3	4	5
21. La homosexualidad es un asunto de las nuevas generaciones	1	2	3	4	5
22. Me preocuparía que un familiar cercano tuviera un maestro homosexual.					
23. Una persona que ha sufrido algún tipo de abuso sexual durante su infancia tiene altas probabilidades de convertirse en homosexual.	1	2	3	4	5
24. Visitaría un *bar gay* en compañía de un amigo o amiga homosexual.	1	2	3	4	5
25. Las personas homosexuales tienen derecho a formar parte de grupos religiosos.	1	2	3	4	5
26. Las parejas homosexuales pueden formar su propio núcleo familiar, siempre y cuando no tengan hijos.	1	2	3	4	5
27. La homosexualidad es producto de un problema genético.	1	2	3	4	5
28. Me preocuparía que un familiar cercano tuviera una maestra lesbiana.	1	2	3	4	5
29. La homosexualidad es producto del entorno social	1	2	3	4	5
30. La verdad es que me divierten los chistes sobre homosexuales.	1	2	3	4	5
31. Si fuera gerente de una empresa le daría trabajo a una persona homosexual.	1	2	3	4	5
32. Me incomodaría tener compañeros de apartamento homosexuales.	1	2	3	4	5
33. La verdad es que la mayoría de los hombres homosexuales son muy afeminados.	1	2	3	4	5
34. Aceptaría un trabajo sin problemas aunque mi jefe inmediato fuera homosexual.	1	2	3	4	5
35. Me siento incómodo (a) cuando veo a dos hombres darse un beso en boca.	1	2	3	4	5
36. Me siento incómodo (a) cuando veo a dos mujeres darse un beso en boca.	1	2	3	4	5
37. Es muy probable que un niño o niña criado por personas homosexuales se vuelva homosexual.	1	2	3	4	5
38. Sería mejor que las personas homosexuales evitaran expresar su orientación sexual en público.	1	2	3	4	5