Departament de Personalitat, Avaluació i Tractaments Psicològics
Facultat de Psicologia
Universitat de Barcelona

Programa de Doctorat en Psicología Clínica i de la Salut
Bienni 2000-2002

Efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios de personalidad

Doctorando **:** Uwe Kramp Denegri

Director de Tesis **:** Dr. Alberto Maydeu Olivares

Barcelona

Marzo de 2006

Dedico este trabajo a mis seres más queridos,

con especial énfasis a Claudia y Tobías.

Claudia, amiga y compañera, gracias por tu paciencia y ayuda.

Sin tu constante apoyo este trabajo nunca hubiese existido.

Tobías, gracias por iluminar con tu sonrisa

cada una de las líneas que dibujan este trabajo.

Agradecimientos

Son numerosas las personas que han ayudado, de forma directa o indirecta, en la realización de este trabajo de Tesis de Doctorado. En primer lugar, quisiera agradecer la valiosa colaboración de todos y cada uno de los *sujetos que cooperaron voluntariamente como muestra*. Sin su participación este trabajo hubiese sido imposible. Asimismo, quisiera agradecer la ayuda prestada por los profesores Dr. *José Gutiérrez Maldonado*, Dr. *Lluis Folch Soler*, Dra. *Consol Marcet Cabral*, Dra. *Alejandra Caqueo Urizar* y la Sra. *Marta Ferre García*, quienes cedieron parte de su valioso tiempo para poder solicitar la colaboración de los participantes en este estudio.

Agradecer también al Dr. *Joan Ferrando Piera*, por su orientación en los estadios iniciales de este trabajo. Como no, agradecer a los Drs. *Josep María Tous Ral* y *Antonio Andrés Pueyo* por sus siempre oportunos comentarios, lo mismo que al Sr. *David Gallardo Pujol* y la Sra. *Donna Coffman*.

Mención especial requiere la colaboración y paciencia prestada por el Sr. *Carlos García Forero* en numerosos apartados técnicos de este trabajo y, por su puesto, agradecer al Dr. *Alberto Maydeu Olivares*, quien, más que un director de tesis, ha sido en todo momento un gran maestro y amigo.

No puedo dejar de mencionar a mi familia, tanto sanguínea como política. Muchas gracias *Ariane Denegri Kossack*, *Rosario Morales Villagran* y *Claudio López de la Maza* por vuestro siempre oportuno apoyo. Habéis sido padres y amigos en todo momento.

El mayor de mis afectos y agradecimientos para *Claudia López Morales*, compañera de travesía. Sin tu constante aliento y apoyo esto no hubiese sido posible. Y, a ti *Tobías*: tu sonrisa ilumina cada una de las líneas trazadas en este trabajo.

Finalmente, agradecer la financiación parcial de los estudios realizados en este trabajo de Tesis de Doctorado al *Ministerio de Ciencia y Tecnología* (Beca BS02003-08507, IP: Dr. Alberto Maydeu-Olivares) y a la *Society of Multivariate Experimental Psychology* (SMEP). Adicionalmente, agradezco la confianza depositada por la SMEP en quien suscribe el presente trabajo, por galardonarlo con uno de sus dos premios otorgados durante el año 2005 (*Dissertation Award, 2005*).

Índice

Resumen

¿Cuántas opciones de respuesta hay que utilizar para responder un cuestionario de personalidad? Durante más de ochenta años, numerosas investigaciones previas han examinado este tópico sin lograr alcanzar un consenso. El presente trabajo se aborda esta cuestión de forma exhaustiva.

Método:

Se aplican dos escalas de personalidad [(*Orientación Negativa hacia los Problemas* (***NPO***) del Social Problem Solving Inventory-Revised (SPSI-R) y la escala *Satisfacción con la Vida* (***SWLS***)], cada una a una muestra independiente (*Estudio A*= 746 participantes y *Estudio B*= 426 participantes), variando el número de opciones de respuesta en *dos*, *tres* y *cinco* alternativas. Cada sujeto debe responder a los tres tipos de formato de respuesta dentro de una misma sesión. Para evitar que éstos noten que están contestando el mismo cuestionario en más de una ocasión, cada escala experimental se inserta dentro de una batería de tests. De esta forma, se construyen tres baterías de test (una para cada muestra), contrabalanceando la aparición de los distintos formatos de respuesta y manteniendo constante el resto de las condiciones experimentales. Por otro lado, se utiliza como criterio para evaluar el efecto del número de opciones de respuesta sobre las propiedades psicométricas las escalas NPO y SWLS tanto su fiabilidad (*consistencia interna* y *estabilidad temporal*) como su validez (*evidencia basada en la estructura interna* y *evidencia basada en el grado de relación con otras variables*). Estas se evalúan desde tres modelos psicométricos diferentes: *Teoría Clásica de los Tests* (TCT), *Análisis Factorial de los Ítems* (AFI) y *Teoría de Respuesta a los Ítems* (TRI).

Aporte:

Tres son los principales aportes de este trabajo de Tesis de Doctorado. En *primer* lugar, se aplican dos escalas de personalidad diferentes a muestras independientes, manteniendo constante el resto de las condiciones experimentales. Con ello se espera controlar que los resultados observados son producto del efecto ejercido por los distintos formatos de respuesta utilizados y no por otras variables. En *segundo* lugar, cada sujeto contesta su respectiva escala experimental utilizando los tres formatos de respuesta propuestos. Ello permite disponer de información directa respecto de cómo ha respondido realmente cada sujeto – y no sólo con una estimación indirecta, derivada de la comparación de muestras independientes, como ha sido lo habitual hasta este momento. *Finalmente*, el análisis de las propiedades psicométricas de las escalas experimentales se aborda por primera vez, simultáneamente, desde tres perspectivas diferentes (TCT, AFI y TRI). Esto permite controlar que los resultados observados no son consecuencia inmediata del modelo psicométrico utilizado, sino producto del efecto ejercido por los distintos formatos de respuesta propuestos. El análisis de la validez de los cuestionarios se realiza aplicando *Modelos de Ecuaciones Estructurales* (SEM).

Resultados y discusión:

En su conjunto, los resultados observados sugieren que el número de opciones de respuesta afecta de forma leve a moderada la *estructura interna* y, parcialmente, la *consistencia interna* de las escalas utilizadas en este trabajo de Tesis de Doctorado. Por el contrario, el número de alternativas de respuesta no afecta aspectos tales como la *evidencia basada en la relación con otras variables* o la *estabilidad temporal* (estabilidad test-retest) de las escalas analizadas.

Palabras clave:

Número óptimo de opciones de respuesta, Cuestionarios de personalidad, Satisfaction with Life Scale (SWLS), Social Problem-Solving Inventory-Revised (SPSI-R), Teoría Clásica de los Tests (TCT), Análisis Factorial de los Ítems (AFI), Teoría de Respuesta a los Ítems (TRI), Modelos de Ecuaciones Estructurales (SEM), Propiedades psicométricas, Fiabilidad, Validez, evidencia basada en la estructura interna, evidencia basada en la relación con otras variables, consistencia interna, estabilidad temporal.

Abstract

Numerous studies have attempted to answer the question of what is the optimal number of response options in personality questionnaires. Some authors argue that the number of response options do not affect at all the reliability or the validity of the measurement tool, whereas others suggest to use from 2- to 25-response alternatives in order to maximize the psychometric properties of the instrument. The present PhD Dissertation work attempts to aboard in an exhaustive manner this question.

Method:

All subjects have responded, within the same session, a questionnaire with *two-*, *three-* and *five-*response options. In order to avoid that the subjects notice that they are responding to the same questionnaire, within a same session, these are applied within a tests battery. Likewise, two independent samples (A and B) and two personality questionnaires [*Negative Problem Orientation* (***NPO***) of the Social Problem Solving Inventory-Revised (SPSI-R) and *Satisfaction with Life Scale* (***SWLS***)] are used, to ensure that the answer does not depend on the instrument investigated. Both, the reliability and validity of the instrument will be used as criteria to determine the optimum number of response alternatives. Finally, the reliability and validity of each response format is analysed using three different psychometric models: *Classical Test Theory* (CTT), *Item Factor Analysis* (IFA), and *Item Response Theory* (IRT). Add, we used *Structural Equation Models* (SEM) to analyze the validity of the NPO and SWLS scales.

Results and discussion:

As a whole, the observed results suggest that the number of response options affects form slight to moderate aspects such as the *internal structure* and,

partially, the *internal consistency* of the scales analyzed in this work. Contrary, the number of response alternatives does not affect aspects such as the *evidence based on the relation with other variables* or the *temporary stability* (test-retest stability) of the examined scales.

Keywords:

Optimal number of response options, Classical Tests Theory (CCT), Item Factor Analysis (IFA), Item Response Theory (IRT), Structural Equation Models (SEM), Psychometric properties, Reliability, Validity, Personality rating scales, Satisfaction with Life Scale (SWLS), Social Problem-Solving Inventory-Revised (SPSI-R), evidence based on the internal structure, evidence based on the relation with other variables, internal consistency, temporary stability.

Prólogo

Este trabajo de Tesis de Doctorado representa la culminación de un camino que comienza el año académico 2000-2001, momento en el cual inicio el Programa de Doctorado *Psicología Clínica y de la Salud*, en el Departamento de Personalidad, Evaluación y Tratamientos Psicológicos (Facultad de Psicología), de la *Universidad de Barcelona*.

Inserto dentro de ámbito general de la *Evaluación Psicológica*, nuestro interés se centra en conocer el ***efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios personalidad.*** Dadas las características del tema en cuestión, se profundizan aspectos relacionados con la *Teoría de los Test*, desde un punto de vista aplicado.

En el entendido que los *cuestionarios de personalidad* son quizás la herramienta de evaluación más difundida dentro del ámbito de la psicología (Fernández-Ballesteros, 2004b; Pelechano, 2000), llama la atención observar como sus constructores rara vez, si acaso alguna, justifican la elección del formato de respuesta utilizado. Es más, las distintas investigaciones empíricas o simulaciones revisadas tampoco parecen resolver esta cuestión (Aiken, 1983; Cox, 1980; Chang, 1994; Churchill y Peter, 1984; Ferrando, 1995; Peabody, 1962; Preston y Colman, 2000; Symonds, 1924). Por lo tanto, la pregunta acerca de cuál debe ser el número óptimo de opciones de respuesta en pruebas que evalúan la personalidad continúa abierta. El presente trabajo aborda de forma exhaustiva esta cuestión.

Lo que sigue se inicia con una **introducción**, donde se dan a conocer los objetivos generales y específicos del presente trabajo de Tesis de Doctorado, así

como los límites dentro de los cuales este se mueve. El resto del trabajo se estructura en tres partes diferentes, complementarias entre sí. La **Parte I** presenta en tres capítulos el *marco teórico* propuesto. Así, el ***capítulo uno*** profundiza aspectos relativos a la definición y descripción de lo que aquí se entiende por *medición de atributos psicológicos*. El ***capítulo dos*** introduce de forma general los principales *modelos de medición* en *Teoría de los Tests*. Se tratan en este capítulo sólo aquellos aspectos que guardan directa relación con los objetivos del presente trabajo. La Parte I finaliza con el ***capítulo tres***, donde se presenta el *estado actual de la cuestión* respecto del estudio del efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios diseñados para la evaluación de las *actitudes* y la *personalidad*.

La **parte II** presenta la metodología de investigación utilizada y expone los resultados obtenidos tras el análisis de datos realizado. Se estructura en dos capítulos. Así, el ***capítulo cuatro*** describe la muestra utilizada (*participantes*), los *instrumentos* aplicados, los *procedimientos* realizados y los *análisis estadísticos* propuestos. Por su parte, el ***capítulo cinco*** presenta los *resultados* observados. Estos se describen y analizan desde la perspectiva de tres modelos psicométricos diferentes: *Teoría Clásica de los Tests* (TCT), *Análisis Factorial de los Ítems* (AFI) y *Teoría de Respuesta a los Ítems* (TRI). Adicionalmente, el análisis de la validez de las escalas de personalidad bajo estudio se realiza aplicando *Modelos de Ecuaciones Estructurales* (SEM).

La **Parte III** cierra este trabajo. Se corresponde con los ***capítulos seis y siete***, donde se ofrece una *discusión* (capítulo 6) *y conclusiones* (capítulo 7) acerca de las implicaciones de los resultados observados y las limitaciones del presente trabajo de Tesis de Doctorado. Asimismo, a lo largo de ambos capítulos, se sugieren nuevas líneas de investigación sobre aspectos aún no resueltos.

Introducción

El presente trabajo se enmarca dentro del ámbito general de la ***evaluación psicológica*** y aplica, de forma particular, aspectos relacionados con la ***teoría de los test***. Específicamente, nuestra intención consiste en *analizar en que medida el número de opciones de respuesta afecta las propiedades psicométricas de los cuestionarios diseñados para la evaluación de la personalidad*.

En un sentido amplio, todo proceso de ***evaluación psicológica*** contempla distintas fases, las cuales pueden orientarse a la *descripción-explicación* de un problema y/o a la *valoración de un tratamiento* o *intervención* efectuada (Fernández-Ballesteros, 2004a). En ambos casos, resulta fundamental contar con medios que permitan recoger información pertinente a los fines de la evaluación. Estos contemplan el uso de distintas técnicas, tales como la *observación*, las *técnicas objetivas* (aparatos e instrumentos), los *autoinformes*, las *entrevistas*, las *técnicas proyectivas*, las *técnicas psicométricas* (tests) o una combinación de las mismas (Fernández-Ballesteros, 2004a; Murphy y Davidshofer, 1994; Pelechano, 2000; Walsh y Betz, 1995). No es el momento ni el lugar para definir o profundizar cada una de las técnicas de evaluación antes mencionadas[1], pero si para indicar que estas herramientas se utilizan con el fin de indagar y cuantificar "…los comportamientos, características o propiedades del sujeto (o sujetos) en estudio y de su contexto" (Fernández-Ballesteros, 2004b, p. 122). El presente trabajo de investigación se centra en uno de estos procedimientos de recogida de datos, como son las *técnicas psicométricas* o *Tests*. Específicamente, en los *cuestionarios* diseñados para la evaluación de la *personalidad*.

[1] Para una revisión detallada de cada una de las técnicas de evaluación mencionadas ver, por ejemplo, Fernández-Ballesteros (2004).

En cuanto al significado del concepto **test**, Renom (1992) indica que estos representan "un conjunto homogéneo y estandarizado de ítems cuyo objetivo es la evaluación cuantitativa bajo condiciones rigurosamente estandarizadas de rasgos y atributos psicológicos y educacionales" (Renom, 1992, p. 581). Por su parte, Cronbach (1990) sugiere que los test se refieren a un "…procedimiento sistemático para observar la conducta y describirla con la ayuda de escalas numéricas o categorías establecidas […] el evaluador recoge información preguntando y observando a todas las personas de la misma manera, en la misma o en comparables situaciones. Esta definición engloba los cuestionarios, a través de los cuales se obtienen informes sobre la personalidad, procedimientos para observar la conducta social, aparatos de medida de la coordinación o, incluso, registros sobre productos" (Cronbach, 1990, p. 32). Martínez-Arias (1996) acota indicando que un test implica un "… reactivo que, aplicado a un sujeto, revela y da testimonio del tipo o grado de su aptitud, de su forma de ser o del grado de instrucción que posee. Estos reactivos o tests constan de preguntas, tareas, estímulos, situaciones, etc. que intentan poner de relieve una muestra de las conductas del sujeto, representativa de la característica que se quiere apreciar o medir (Martínez-Arias, 1996, p. 31). Por su parte, los *Standards for Education and Psychological Testing* (de aquí en adelante *Standards*) indican que todo test se refiere a un sistema diseñado para recolectar ejemplos del trabajo o conductas particulares de un individuo, en un área en particular (American Psychological Association, American Educational Research Association y National Council on Measurement in Education, 1999).

En definitiva, los test psicológicos representan, parafraseando a Anastasi (1982), *una medida objetiva y estandarizada de una muestra de conducta*. Como se desprende de las definiciones anteriores, esta *medida* puede ser sobre (a) el *rendimiento*, tanto académico como profesional, (b) la *inteligencia* o las *aptitudes* y (c) la *personalidad*, las *actitudes* o los *intereses* de una persona o grupo de personas

(Martínez-Arias, 1996). Como ya se adelantó, este trabajo centra su atención en el tercero de estos bloques de medidas. Específicamente, sobre los cuestionarios diseñados para la evaluación de la personalidad.

En términos amplios, un **cuestionario** de personalidad es una herramienta compuesta por una serie de frases o preguntas que hacen referencia a aspectos afectivos, cognitivos o comportamentales que deben ser evaluados como pertenecientes o no al ámbito conductual personal (autoevaluación) o de un tercero (heteroevaluación). En otras palabras, implican una auto- o heteroevaluación que contiene preguntas estructuradas destinadas a evaluar la frecuencia o intensidad en que se manifiesta determinado atributo o rasgo. "Como su nombre lo indica, estos tests consisten en una serie de ítems, que suelen ser frases en formas variadas, tales como preguntas, preposiciones o afirmaciones que están relacionadas con los comportamientos, sentimientos, pensamientos o creencias más habituales y frecuentes entre las personas. A los sujetos se les solicita que contesten de forma apropiada a los citados ítems […] En general, los cuestionarios pretenden obtener información de las tendencias de respuesta y conducta de los individuos a través de su información (autoevaluación) o de la información aportada por personas cercanas o conocedoras de éstos (heteroevaluación) (Andrés, 1997, pp. 455-456). Su aplicación puede ser individual o colectiva. Dada la sencillez y lo económico de su aplicación, son consideradas como la herramienta de evaluación más difundida dentro del ámbito de la psicología (Fernández-Ballesteros, 2004b; Pelechano, 2000). Por otro lado, aunque no exclusivamente, el modelo de los rasgos ha servido históricamente como base para la formulación y desarrollo de este tipo de tests (Calero y Padilla, 2004; Walsh y Betz, 1995).

Tanto para el caso de los tests en general, como para el de los cuestionarios en particular, la ***Teoría de los Tests*** se ocupa de construir modelos orienta-

dos a explicar y evaluar las características psicométricas de tales instrumentos de medición. En un sentido amplio, la Teoría de los Tests se ocupa de estudiar los factores que inciden sobre las puntuaciones observadas en los tests, formulando para ello una serie de modelos explicativos. Entre otros, la *Teoría Clásica de los Tests* (TCT), el *Análisis Factorial de los Ítems* (AFI) y la *Teoría de Respuesta a los Ítems* (TRI) suelen ser reconocidos como los modelos psicométricos más difundidos (Allen y Yen, 1979; Gulliksen, 1987; Hambleton y Swaminathan, 1985; Lord, 1980; Martínez-Arias, 1996; McDonald, 1999; Muñiz, 1996; Santisteban-Requena, 1990). La presente investigación se centra en el análisis del efecto que puede tener el número de opciones de respuesta sobre la fiabilidad y validez de los cuestionarios de personalidad, desde la perspectiva de los tres modelos psicométricos antes señalados (TCT, AFI y TRI). Cabe advertir que la elección de estas tres perspectivas se basa en la intención de abarcar de la forma más exhaustiva posible la cuestión planteada. Nuestro fin es puramente pragmático, por lo que no se entra en un análisis comparativo acerca de las bondades o dificultades que pueda presentar cada uno de los modelos psicométricos mencionados.

En lo que se refiere al estudio del *efecto del número de opciones de respuesta sobre las propiedades psicométricas de los instrumentos de evaluación psicológica* propiamente tal, hay que señalar que el tema no es nuevo. Los primeros trabajos se remontan hasta el primer tercio del siglo XX y se centran principalmente en pruebas diseñadas para evaluar el ***rendimiento académico***. De acuerdo a Lord (1977, 1980), los primeros en investigar este tema son Toops (1921), Ruch y Stoddard (1925), Ruch, DeGraff y Gordon (1926), Ruch y Stoddard (1927) y Ruch y Charles (1928). En general, estos trabajos coinciden en señalar que ***tres*** es el número óptimo de alternativas de respuesta, en tanto permite garantizar una adecuada ***fiabilidad*** en los instrumentos que evalúan el rendimiento académico. Estudios empíricos posteriores parecen confirmar esta conclusión (Cos-

tin, 1970, 1972; Rogers y Harley, 1999; Straton y Catts, 1980; Tversky, 1964), la cual cuenta con respaldo tanto desde la perspectiva de la *Teoría Clásica de los Tests* (Bruno y Dirkzwager, 1995; Costin, 1970, 1972; Garner, 1960; Grier, 1975, 1976; Rogers y Harley, 1999; Straton y Catts, 1980; Tversky, 1964) como desde la *Teoría de Respuesta a los Ítems* (Abad, Olea y Ponsoda, 2001; Lord, 1977, 1980). En la misma línea, Rogers y Harley (1999) señalan que para este tipo de instrumentos, "trabajos teóricos y simulaciones de tests (…) revelan que tres opciones, en tests con ítems de respuesta múltiple, son tan buenos como los ítems de respuesta múltiple con cuatro o cinco opciones de respuesta en términos de discriminación del ítem, a nivel de los ítems, y respecto de la consistencia interna y el error estándar de medida, a nivel del test" (Rogers y Harley, 1999, p. 236).

Contrariamente, dentro del ámbito de la ***evaluación de las actitudes y de la personalidad*** aún no queda claro cuál debe de ser el número óptimo de alternativas de respuesta (Bearden, Netmeyer y Mobley, 1993; Cox, 1980; Peter, 1979; Shaw y Wright, 1967). Si se analizan las investigaciones realizadas en los últimos ochenta años, tanto empíricas como simulaciones, la afirmación anterior no parece sorprendente. Mientras que algunos autores sostienen que basta con utilizar *dos* o *tres* alternativas de respuesta (Aiken, 1983; Komorita y Graham, 1965; Masters, 1974; Matell y Jacoby, 1971, 1972; Peabody, 1962; Sancerni, Meliá y González, 1990), otros sugieren utilizar desde *cinco* (Churchill y Peter, 1984; García-Cueto, Muñiz y Lozano, 2002) o *siete* opciones (Ferrando, 1995; Preston y Colman, 2000; Ramsay, 1973; Symonds, 1924), hasta un máximo de *25* alternativas de respuesta (Champney y Marshall, 1939) para maximizar las propiedad psicométricas del instrumento. Es más, si se tuviese que establecer democráticamente el número óptimo de alternativas, *cinco* parece ser el número de opciones más utilizado, seguido por *siete* categorías de respuesta (Cox, 1980; Churchill y Peter, 1984; Ferrando, 1995; Preston y Colman, 2000; Ramsay, 1973).

Frente a tal disparidad de conclusiones, surge necesariamente la pregunta acerca de ***por qué*** esta falta de acuerdo. A nuestro juicio, tres son las razones principales. La ***primera*** se relaciona con el método de recogida de datos. Algunos investigadores tienden a recoger un número elevado de alternativas de respuestas y luego *colapsan* las mismas en distintos formatos de respuesta arbitrarios (p.ej., Matell y Jacoby, 1971; Peabody, 1962). Otros investigadores prefieren administrar las distintas condiciones experimentales propuestas en muestras independientes, las cuales comparan luego entre sí (p. ej., Chang, 1994; Masters, 1974; Preston y Colman, 2000). Este procedimiento, aunque útil, presenta una dificultad importante: no permite recoger información de forma directa respecto de *cómo* ha respondido realmente el sujeto, en el entendido que solo simula este hecho. La ***segunda*** razón se refiere a las discrepancias respecto de *qué criterio* utilizar como mejor indicador para establecer un número óptimo de alternativas de respuesta. En general, la mayoría de los investigadores suele elegir la *fiabilidad* como criterio (p.ej., Aiken, 1983; Bandalos y Enders, 1996; Cicchetti, Showalter y Tyrer, 1985; Jenkins y Taber, 1977; Komorita y Graham, 1965; Lissitz y Green, 1975; Masters, 1974; Weng, 2004). Otros, aunque en menor medida, recurren a la *validez* (p.ej., Comrey y Montag, 1982; Olsson, 1979) o simultáneamente a éstos dos criterios (p.ej., Chang, 1994; García-Cueto *y col.*, 2002; Matell y Jacoby, 1971, 1972; McCallum, Keith y Wiebe, 1988; Preston y Colman, 2000; Sancerni *y col.*, 1990). Finalmente, una ***tercera*** razón se relaciona con el hecho de que este problema puede ser abordado desde diferentes *modelos psicométricos*, tales como la *Teoría Clásica de los Test* (TCT) (p. ej., Aiken, 1983; Bandalos y Enders, 1996; Cox, 1980; Masters, 1974; Weng, 2004), el *Análisis Factorial de los Ítems* (AFI) (p. ej., Comrey y Montag, 1982; Chang, 1994; García-Cueto *y col.*, 2002; McCallum *y col.*, 1988; Olsson, 1979) o la *Teoría de Respuesta a los Ítems* (TRI) (p. ej., García-Cueto, Muñiz y Lozano, 2003; Hernández, Muñiz y García-Cueto, 2000).

Ante este panorama, la presente investigación pretende abordar de forma exhaustiva el análisis del efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios de personalidad. Para ello, se realizaron dos estudios, identificados como *Estudio A* y *Estudio B*. En cada estudio se trabajó con una muestra empírica independiente, a la que se aplicó una misma batería de tests, variando sólo lo que hemos denominado como *escala experimental*. Así, el ***Estudio A*** trabajó experimentalmente con la escala *Orientación Negativa hacia los Problemas* (**NPO**, D'Zurilla, Nezu y Maydeu-Olivares, 2002), mientras que en el ***Estudio B*** se aplicó la escala *Satisfacción con la Vida* (**SWLS**, Diener, Emmons, Larsen y Griffin, 1985). El resto de los cuestionarios que componen cada batería de tests se mantuvieron constantes en ambos estudios; a saber, el *Inventario NEO de los Cinco Factores* (*NEO-FFI*, Costa y McCrae, 1992, 1999), el *Inventario de Afectos Positivos y Afectos Negativos* (*PANAS*, Watson, Clark y Tellegen, 1988) y el *Inventario Revisado de Resolución de Problemas Sociales* (*SPSI-R*, D'Zurilla *y col*., 2002). La inclusión de estos últimos cuestionarios cumple dos objetivos: (1) servir de distractores y (2) actuar como variables externas (variables criterio) para el análisis de la validez de las escalas experimentales (NPO y SWLS).

Siguiendo con la presentación del método de investigación utilizado, en ambas escalas experimentales se modificó el formato de respuesta variando el número de opciones de respuesta en *dos*-, *tres*- y *cinco*-alternativas. Dentro de una misma sesión, cada sujeto respondió la escala NPO (Estudio A) o la escala SWLS (Estudio B) utilizando los tres formatos de respuesta propuestos. Pesamos que esta estrategia permite recabar información directa respecto de cómo han contestado realmente los participantes.

En cuanto a los criterios utilizados para definir un número óptimo de opciones de respuesta, se recurre tanto a la fiabilidad del instrumento [*consistencia*

interna y *estabilidad temporal* (American Psychological Association *y col.*, 1999)] como a su validez [*evidencia basada en la estructura interna* y *evidencia basada en el grado de relación con otras variables* (American Psychological Association *y col.*, 1999)]. Finalmente, el análisis de los datos se llevó a cabo desde tres modelos psicométricos diferentes: *Teoría Clásica de los Tests* (TCT), el *Análisis Factorial de los Ítems* (AFI) y la *Teoría de Respuesta a los Ítems* (TRI). En cada caso, se analiza la validez de las escalas experimentales por medio de *Modelos de Ecuaciones Estructurales* (SEM).

En función del método de investigación propuesto, esperamos poder encontrar una solución única, independientemente de la escala experimental o del modelo psicométrico utilizado, respecto del efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios de personalidad. Confiamos en que este trabajo permita establecer las bases para cerrar el debate sobre el tema en cuestión o, en su defecto, estimule nuevos estudios que permitan hacerlo, siguiendo o ampliando la línea de investigación propuesta.

Sección I: Marco teórico

En la introducción se han presentado el contexto y los objetivos generales que orientan el presente trabajo de Tesis de Doctorado. A continuación se describe el **marco teórico** sobre el cual descansa este trabajo. En tal sentido, el ***capítulo 1*** aborda lo que aquí se entiende por *medición de atributos psicológicos*. El ***capítulo 2*** describe de forma amplia los *modelos de medición* más difundidos dentro de la Teoría de los Tests. Así, este capítulo se inicia con la presentación del Modelo Teoría Clásica de los Tests (TCT, apartado 2.1). Luego se dan a conocer los modelos de Análisis Factorial de los Ítems (AFI, apartado 2.2) y de Teoría de Respuesta a los Ítems (TRI, apartado 2.3). Finalmente, el ***capítulo 2*** se cierra con una descripción general del Modelo de Ecuaciones Estructurales (SEM, apartado 2.4), en tanto éste último sirve de base para estimar la validez de las escalas de personalidad analizadas en la sección II, desde la perspectiva de los modelos de TCT, AFI y TRI. Por otro lado, el ***capítulo 3*** expone el *estado de la cuestión* respecto del tema tratado en este trabajo.

1. *Medición de atributos psicológicos*

Toda empresa científica cuenta con un paradigma, teoría, modelo o, al menos, un esquema racional básico que la sustenta. En muchos casos, éstos no se hacen explícitos, aunque sí los aspectos metodológicos (instrumentales y pragmáticos) que permiten llevar a cabo investigaciones basadas en los mismos. Dentro del ámbito del estudio de la personalidad en particular y de la psicología en general, algunos autores llegan incluso a sostener que el énfasis puesto en aspectos relativos a la evaluación por sobre el desarrollo de teorías básicas, conduce a la muerte de estas últimas (Royce y Powell, 1983). Sin querer entrar más allá de lo necesario en esta polémica, pensamos que la posición anterior es extrema y algo alejada de la realidad. Creemos que la escisión entre teoría y método es más bien aparente, al tiempo que estamos de acuerdo con Pelechano (2000) cuando indica que una sin la otra conduce a un *espacio vacío*. "…la metodología, tanto en el sentido de procedimiento como de teoría de la evaluación, representa un elemento importante en el funcionamiento y delimitación conceptual de lo que significa ‹‹científico››, pero […] la ciencia no se agota en él y, en especial, el método es mucho más importante en el contexto de la justificación que en el del descubrimiento. Y no debería de olvidarse que el primero, sin el segundo se convierte muy rápidamente en una suerte de espacio ‹‹vacío››" (Pelechano, 2000, p. 79).

Coherentes con lo expuesto hasta el momento, el presente trabajo aborda ciertas preguntas respecto del diseño y uso de los cuestionarios de personalidad que aún no han encontrado una respuesta definitiva ¿El número de opciones de respuesta tiene efectos sobre las propiedades psicométricas de los cuestionarios de personalidad? De ser así, ¿en qué medida? y ¿bajo que condiciones? Antes de tratar estas cuestiones en profundidad, creemos conveniente hacer explícito el

esquema racional sobre el cual se fundamenta este trabajo de Tesis de Doctorado. En tal sentido, a continuación se presenta nuestra posición respecto de lo que entendemos por *medición en psicología*, al tiempo que se indican las formas habituales de resolver esta cuestión, desde el ámbito de la *psicometría*.

Entendemos por ***medición*** el acto por medio del cual se asignan números a los atributos de un objeto que se desea evaluar, siguiendo ciertas reglas previamente definidas por la comunidad científica. Para Stevens (1951) "medir es la asignación de números a los objetos o eventos de acuerdo a ciertas reglas" (Stevens, 1951, p. 22), mientras que para Campbell (1938), la medición consiste en "la asignación de números para representar las propiedades presentes en los sistemas materiales, distintos de los números, en virtud de ciertas reglas que gobiernan tales propiedades" (Campbell, 1938, p. 126). La *Teoría de la Medición* es la responsable de definir las reglas que permiten asignar números a las propiedades de los objetos. Se han desarrollado distintas propuestas para abordar este tópico (p. ej., Coombs, 1964; Luce y Tukey, 1964; Torgerson, 1962), a pesar de lo cual la *Teoría Representacional de la Medición* (Stevens, 1946, 1951), aunque no exenta de críticas, sigue siendo el modelo más difundida dentro del ámbito de la psicología (Andrés, 1997; Fernández-Ballesteros, 2004a; Martínez-Arias, 1996).

Una vez establecidas las reglas que permiten la *asignación de números a los objetos o eventos* (Stevens, 1951), el científico puede preguntarse una cuestión más abstracta y compleja ¿Cómo medir en psicología? Lo que sigue profundiza esta cuestión, adaptando lo que Torgerson (1962) entiende por *estructura básica de las Ciencias Sociales* (ver figura 1).

De acuerdo a lo señalado por Torgerson (1962), para encarar su objeto de estudio, el científico debe establecer en primer lugar un modelo o marco teórico

de referencia. En el caso de la psicología, este se compone de uno o varios constructos psicológicos, operacionalmente definidos, los cuales dan cuenta de las propiedades y descripción de uno o varios atributos[2] presentes en el objeto evaluado. Junto con distinguirlos claramente entre sí, es necesario también hacer explícita la relación que existe entre los constructos que dan cuerpo a la teoría. Una vez definido el marco teórico, es necesario establecer un puente de unión entre la formulación teórica y la realidad empírica. La *medición* se encarga de unir estos ámbitos, utilizando para ello herramientas de medición pertinentes (p. ej., cuestionarios de personalidad). En términos pragmáticos, esta unión se materializa por medio de la asignación sistemática de números a los atributos bajo estudio (p. ej., a través de la *Teoría Representacional de la Medición*). La finalidad de este procedimiento no es otra que hacer inferencias válidas acerca del constructo evaluado. En caso de existir correspondencia entre lo evaluado y la formulación teórica, la comunidad científica puede aceptar provisionalmente esta última. La figura 1 intenta reflejar de modo gráfico lo señalado hasta el momento.

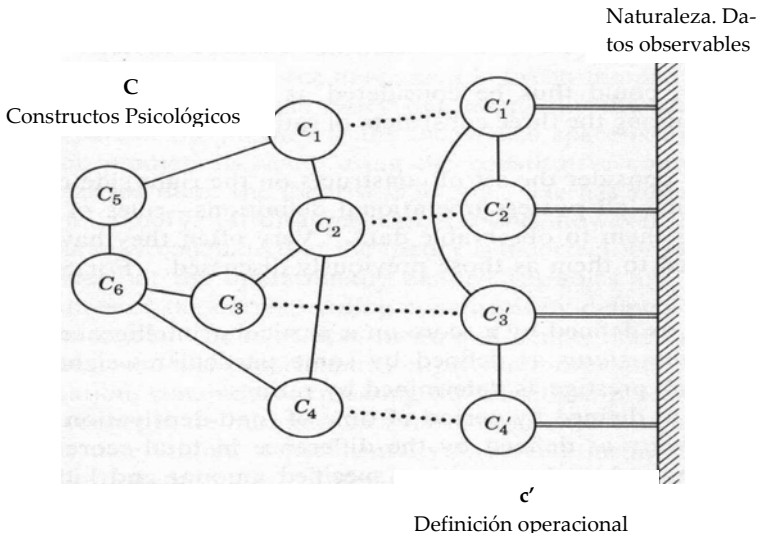


Figura 1: Estructura básica de las Ciencias Sociales. Adaptado de Torgerson (1962).

[2] Torgerson (1962) entiende por ***atributo*** aquellas *características* que dan cuenta del objeto de estudio, las cuales permiten definir y describir el mismo.

De acuerdo a Torgerson (1962), los círculos que aparecen a la izquierda de la figura 1 (C_1, C_2…C_n) representan los distintos constructos o atributos psicológicos que dan cuerpo al modelo o teoría. Su formulación es semántica y se refieren a aquellos conceptos *conductual y socialmente importantes* que intenta explicar el investigador. Por su parte, la relación entre los distintos constructos queda reflejada por las líneas continuas simples que aparecen a la izquierda de la figura 1. Por ejemplo, alguien puede estar interesado en formular un modelo que explique el *bienestar psicológico subjetivo* (Diener *y col.*, 1985). De acuerdo con lo señalado hasta el momento, en primer lugar se debe proponer una definición *semántica* que describa claramente lo que se entiende por bienestar psicológico subjetivo. Luego, es necesario definir, también semánticamente, los elementos que hipotéticamente componen este último. Así, el investigador puede indicar que el bienestar psicológico subjetivo es una dimensión de personalidad que surge de la relación entre tres elementos: afectos positivos (AP), afectos negativos (AN) y satisfacción con la vida (SV) (Diener *y col.*, 1985).

Volviendo a la figura 1, para validar o refutar científicamente un modelo o teoría, es fundamental relacionarlo de manera formal con la naturaleza. De lo contrario solo se cuenta con una especulación filosófica. La definición operacional y formal de los constructos, tratados originalmente de forma semántica, permite establecer un puente entre la teoría y la realidad empírica. Lo anterior está representado en la figura 1 por medio de los círculos C'_1, C'_2…C'_n, en tanto que las líneas discontinuas intentan reflejar el paso de una definición semántica a otra de tipo formal. Por regla general, los constructos definidos tanto semántica como formalmente suelen recibir el mismo nombre.

Siguiendo con nuestro ejemplo, el investigador puede señalar que el bienestar psicológico subjetivo es una dimensión de personalidad que surge de la relación que se establece entre AP, AN y SV. Por lo tanto, es fundamental de-

finir operacionalmente tales componentes. Así, puede indicar que AP refleja el grado en que cada persona siente entusiasmo, energía o alerta (por ejemplo, energía, concentración, interés, etc.) (Watson *y col.*, 1988). De AN puede decir que representa una dimensión general ligada a la experimentación subjetiva de sensaciones de angustia o poco placenteras, la cual agrupa estados de ánimo diversos, tales como ira, desprecio, disgusto, etc. (Watson *y col.*, 1988). De SV puede decir que representa la percepción subjetiva sobre los asuntos de la vida diaria, la cual está basada en la comparación con los estándares personales de cada individuo, no externamente impuestos (Diener *y col.*, 1985). En cada caso, es necesario establecer las conductas socialmente relevantes o atributos que dan cuenta de cada definición dada. Finalmente, puede definir formalmente el concepto de bienestar psicológico subjetivo, descrito previamente como una dimensión de personalidad, como la puntuación obtenida por el sujeto tras contestar un cuestionario o batería de tests de auto-informe compuesta por reactivos que evalúan conductas relacionadas con AP, AN y SV (Diener *y col.*, 1985).

Es importante tener presente que los constructos definidos operacional y formalmente son sólo un *indicador* de los constructos que aparecen más a la derecha de la figura 1. "En el mejor de los casos, se *presume* que ambos están monotónicamente relacionados. En el peor de los casos, se *presume* que existe al menos una correlación positiva entre estos" (Torgerson, 1962, p. 7). Las líneas discontinuas de la figura 1 representan esta presunta relación monotónica.

Todavía queda otro punto por resolver, como es el establecimiento de reglas de correspondencia entre los conceptos teóricamente formulados y su representación empírica; es decir, cómo medir. Estas reglas permiten, entre otras cosas, construir herramientas pertinentes capaces de justificar o refutar la teoría o modelo previamente formulado.

Volviendo a la figura 1, el puente de unión entre la teoría y la realidad empírica queda reflejado por las líneas dobles que aparecen a la derecha de la misma. En términos prácticos, estas líneas reflejan el acto de medir, el cual ya ha sido definido anteriormente. Para hacer operativo dicho acto, en psicología se cuenta con distintas técnicas de evaluación, tales como la observación, las técnicas objetivas (instrumentos y aparatos), los autoinformes, las entrevistas, las técnicas proyectivas y las técnicas psicométricas (tests) o una combinación de los mismos (Fernández-Ballesteros, 2004a; Murphy y Davidshofer, 1994; Walsh y Betz, 1995).

Siguiendo con nuestro ejemplo, se ha dicho que el bienestar psicológico subjetivo queda definido operacionalmente por medio de la puntuación obtenida por el sujeto tras responder un cuestionario (test) o batería de tests. En tanto dimensión de personalidad y de acuerdo al modelo propuesto por Stevens (1946, 1951), las propiedades del atributo sugerido en nuestro ejemplo admite una escala de medida de intervalo. Por lo tanto, el instrumento diseñado para su evaluación debe contar con una *escala* que permita (a) reflejar la presencia o ausencia de los atributos que dan cuenta del bienestar psicológico subjetivo (AP, AN y SV), (b) ordenar la magnitud relativa de cada uno de los mismos según la opinión del sujeto evaluado y (c) establecer relaciones de igualdad y diferencia entre las magnitudes de AP, AN y SV, con objeto de establecer la posición relativa del sujeto en cada una de estas dimensiones.

Solucionado el problema acerca de cómo asignar reglas de medición, estamos en condiciones de finalizar la descripción del modelo sobre la *estructura básica de las Ciencias Sociales* propuesto por Torgerson (1962). Como se indicó anteriormente, la definición operacional de los constructos permite establecer un puente de unión entre la teoría o modelo psicológico y la realidad empírica. En la figura 1, esta última está representada por la línea vertical sombreada que

aparece a la derecha (Naturaleza. Datos empíricos). Naturalmente, es ésta última la responsable de aportar los datos relevantes que permiten validar o refutar la teoría o modelo propuesto. Lo anterior se consigue a través del uso de instrumentos de medición fiables y válidos, respecto de las propiedades de los atributos bajo estudio. Como ya se indicó anteriormente, la teoría de los tests se encarga de establecer los modelos psicométricos responsables de guiar la construcción de un instrumento eficiente y eficaz (p. ej., cuestionario), siendo los más difundidos la *Teoría Clásica de los Tests* (TCT), el *Análisis Factorial de los Ítems* (AFI) y la *Teoría de Respuesta a los Ítems* (TRI) (Allen y Yen, 1979; Gulliksen, 1987; Hambleton y Swaminathan, 1985; Lord, 1980; Martínez-Arias, 1996; McDonald, 1999; Muñiz, 1996; Santisteban-Requena, 1990). Éstos se describen en el capítulo siguiente.

2. *Modelos de medición en psicometría*

La mayor parte de las mediciones, sea esta física o psicológica, presenta siempre algún grado de error. Este puede provenir del objeto evaluado, del instrumento utilizado, del evaluador o una combinación de tales fuentes (Fernández-Ballesteros, 2004a; Gulliksen, 1987; Lord y Novick, 1968; Nunnally, 1967). En relación al tema que aquí nos ocupa, la ***Teoría de los Test*** se encarga de establecer *modelos* capaces de evaluar las *propiedades psicométricas* de los tests psicológicos[3]. Específicamente, estudia aquellos factores que influyen sobre la puntuación observada en los tests y propone modelos que permiten estimar la puntuación verdadera obtenida por el sujeto, esto último a partir de inferencias basadas en las puntuaciones observadas (Fernández-Ballesteros, 2004a; Gulliksen, 1987; Hambleton y Swaminathan, 1985; Lord, 1980; Lord y Novick, 1968; Martínez-Arias, 1996; Muñiz, 1996, 1997, 2003; Nunnally, 1967). En la actualidad, tres son los modelos que cuentan con mayor difusión: Teoría Clásica de los Test (TCT), Análisis Factorial de los Ítems (AFI) y Teoría de Respuesta a los Ítems (TRI) (Allen y Yen, 1979; Fernández-Ballesteros, 2004a; Gulliksen, 1987; Hambleton y Swaminathan, 1985; Lord, 1980; Martínez-Arias, 1996; McDonald, 1999; Muñiz, 1996; Santisteban-Requena, 1990).

El presente capítulo aborda los modelos de medición antes señalados, poniendo especial énfasis en su caracterización y propuesta acerca de cómo evaluar las propiedades psicométricas (fiabilidad y validez) de las pruebas psicológicas. A su vez, en función de los supuestos subyacentes a los modelos TCT, AFI y TRI, se introducen también algunos conceptos básicos relacionados con el modo de estimar la validez de los instrumentos de medición por medio de *Modelos de Ecuaciones Estructurales* (SEM).

[3] De aquí en adelante, se utiliza de forma indistinta los conceptos *cuestionario* o *test*.

Naturalmente, presentar de forma exhaustiva cada uno de los modelos antes indicados (TCT, AFI, TRI y SEM) es una tarea que excede los fines de este trabajo. Por lo tanto, se ofrece sólo una descripción general de los mismos, acorde a las necesidades de las investigaciones que aquí se proponen. Para aquellos que deseen profundizar sobre los supuestos y aplicaciones de estos modelos psicométricos, se sugiere consultar las distintas obras especializadas existentes. Entre otras, para la ***TCT***, Gulliksen (1987), Lord y Novick (1968), Martínez-Arias (1996), McDonald (1999) o Muñiz (2003). Para ***IFA***, Lawley y Maxwell (1971), Martínez-Arias (1996), Maydeu-Olivares y McArdle (2005), McDonald (1985, 1999) o Mulaik (1972). Para ***TRI***, Hambleton y Swaminathan (1985), Lord (1980), Martínez-Arias (1996), Maydeu-Olivares y McArdle (2005), McDonald (1999), Muñiz (1997) o Van der Linden y Hambleton (1997). Y para ***SEM***, Bentler (1980), Bollen (1989; Bollen y Long, 1993), Gómez (1996), McDonald (1999) o Maydeu-Olivares y McArdle (2005). Antes de presentar los mismos, se definen los conceptos de *fiabilidad* y *validez*.

Respecto del concepto **propiedades psicométricas**, entendemos por tal el análisis de las características, matemática y estadísticamente evaluadas, que dan cuenta de la idoneidad del instrumento de medida. En un sentido amplio, lo anterior puede ser definido como un *proceso* en el cual se procura resolver dos aspectos fundamentales en todo test, como son su fiabilidad y validez. La ***fiabilidad*** se refiere a la necesidad de evaluar la consistencia las puntuaciones ofrecidas por un test. Desde una perspectiva clásica, ésta se define como "… el coeficiente de correlación entre una forma experimental de un test y una forma hipotéticamente equivalente" (Spearman, 1927). Por su parte, los *Standards* definen la fiabilidad como el proceso a través del cual se comprueba que las medidas ofrecidas por un test son consistentes, tras aplicar el mismo en distintas ocasiones sobre una muestra de individuos o grupos (American Psychological Association *y col.*, 1999). Los *Standards* señalan también que para evaluar la fiabilidad

se pueden utilizar tres métodos de estimación: (a) *coeficiente de correlación*, (b) *modelos de ecuaciones estructurales* (SEM) y (c) *fiabilidad inter-jueces* (American Psychological Association *y col.*, 1999). En lo que a nosotros toca, la presente investigación evalúa el efecto del número de opciones de respuesta sobre la ***fiabilidad*** de los instrumentos experimentales utilizados, tomando como referencia el método basado en el coeficiente de correlación (American Psychological Association *y col.*, 1999). Específicamente, se evalúa la *consistencia interna* y la *estabilidad temporal* de los mismos, según se define y describe más adelante, durante la presentación de cada modelo.

Por otro lado, la ***validez*** se refiere a la necesidad de garantizar que las inferencias derivadas del instrumento de medida sean pertinentes y acordes al atributo que se supone se está evaluando. En otras palabras, se espera que el instrumento mida realmente lo que se dice estar evaluando y no otra cosa. Así, mientras la fiabilidad se centra en el estudio de la consistencia de las puntuaciones observadas, la validez permite garantizar la idoneidad de las inferencias derivadas a partir de las puntuaciones observadas, en el sentido de que éstas se refieran única y exclusivamente al constructo evaluado *(American Psychological Association y col., 1999)*.

Históricamente, la definición de lo que significa validez ha variado sustancialmente. Naturalmente, revisar tal evolución no se corresponde con los propósitos de este trabajo, por lo que se sugiere consultar Fernández-Ballesteros (2004a), Martínez-Arias (1996), Muñiz (1996, 2003) o Wainer y Braun (1988). Sin embargo, es interesante recordar que la comprensión del concepto en cuestión ha variado, desde un concepto puramente pragmático y centrado en el instrumento (relación entre un test y un criterio), hasta un concepto unitario compuesto por distintos aspectos que se relacionan de forma directa con la validación del constructo evaluado y no con el test en sí mismo (Fernández-

Ballesteros, 2004b). En tal sentido, la última edición de los *Standards* define la validez como un concepto *unitario* que da cuenta del "…grado en que todas las evidencias acumuladas apoyan la proyectada interpretación de las puntuaciones del test implicadas en el propósito propuesto […]. Validez se refiere al grado en que evidencias y teoría soportan las interpretaciones de las puntuaciones del test implicadas en los usos propuestos del mismo […]. El proceso de validación implica la acumulación de evidencias que proporcionen una base científica sólida para las interpretaciones que se proponen. Esto es, deben ser evaluadas las interpretaciones de las puntuaciones del test requeridas para usos propuestos, no el test mismo" (American Psychological Association *y col.*, 1999, pp. 9-11).

Tradicionalmente se ha propuesto evaluar tres tipos de validez: de contenido, de criterio y de constructo (Martínez-Arias, 1996; Muñiz, 1996, 2003; Wainer y Barun, 1988). De todos modos, en la actualidad los *Standards* proponen recoger *evidencias* que dan lugar a un proceso de evaluación que puede contemplar la valoración de hasta cinco tipos diferentes de validez: (a) validez de contenido, (b) validez del proceso de respuesta, (c) estructura interna, (d) relación con otras variables y (e) validez de las consecuencias de la aplicación del instrumento (American Psychological Association *y col.*, 1999). En lo que a nosotros toca, la presente investigación analiza el efecto del número de opciones de respuesta sobre la ***validez*** de los instrumentos utilizados, tomando como referencia su influencia sobre la *estructura interna* y la *relación con otras variables* de los mismos. De acuerdo a los *Standards*, la validez relacionada con la ***estructura interna*** (o *evidencia basada en la estructura interna*) alude al análisis de la relación de los ítems entre sí y de éstos con el constructo evaluado para interpretar las puntuaciones ofrecidas por el instrumento (p. ej., análisis factorial o análisis de consistencia). Por su parte, la evidencia basada en la ***relación con otras variables*** (o *evidencia basada en la relación con otras variables*), implica relacionar las

puntuaciones del test con criterios externos, los cuales evalúan el mismo constructo, aspectos relacionados con éste o constructos diferentes (p. ej., análisis de evidencias convergentes y discriminantes, aplicación de modelos de ecuaciones estructurales, etc.).

Dicho lo anterior, estamos en condiciones de presentar los supuestos y características generales de TCT, AFI y TRI, en tanto modelos, como también las soluciones propuestas por cada uno de éstos para estimar las propiedades psicométricas del instrumento de medición. Adicionalmente, se introducen los supuestos y soluciones ofrecidas por los modelos SEM, en tanto éstos últimos son utilizados como base para evaluar la validez (*evidencia basada en la estructura interna* y *evidencia basada en la relación con otras variables*) de los instrumentos analizados en este trabajo.

2.1. Teoría Clásica de los Tests

La *Teoría Clásica de los Tests* (TCT) es el primer modelo propuesto para abordar aspectos relacionados con la fiabilidad del instrumento de medida. Surge a comienzos del siglo XX, con la formulación que hace Spearman (1904a, 1904b, 1907, 1910, 1913, 1927) de lo que hoy se conoce como *Modelo de la Puntuación Verdadera* o TCT (Martínez-Arias, 1996; McDonald, 1999). En 1950 Gulliksen (1987) escribe el libro *Theory of Mental Tests*, donde sintetiza lo conocido hasta ese momento en TCT. Posteriormente, Lord y Novick (1968) publican el libro *Statistical Theories of Mental Tests Scores*, en el que se plantea una revisión crítica de la TCT, al tiempo que proponen nuevas líneas de trabajo, entre las cuales se encuentra la Teoría de Respuesta a los Ítems (TRI). Naturalmente, son muchas más las obras publicadas acerca de la TCT. Se han mencionado solo éstas, en tanto han sido calificadas por algunos como las más relevantes sobre la materia

(Muñiz, 1996, 2003). Lo que sigue es, en buena media, una adaptación de Gulliksen (1987), la cual se ve complementada con los aportes realizados por autores contemporáneos.

En cuanto a su formulación general, la TCT propone un *modelo lineal* en el que se asume que la puntuación obtenida por el sujeto *i* en un test (*X* o *puntuación empírica*) se compone de dos elementos aditivos: (1) la *puntuación verdadera* (*T*) obtenida por el sujeto y (2) *el error de medida* (*e*) presente en las puntuaciones observadas. Formalmente, lo anterior queda definido por la ecuación (2.1).

$$X_i = T_i + e_i \tag{2.1}$$

De acuerdo a los señalado por Allen y Yen (1979), el modelo TCT asume que: (a) la correlación entre la puntuación verdadera (*T*) y los errores de medida (*e*) es igual a cero; (b) los errores de medida (*e*) para formas paralelas no están correlacionados entre sí; (c) los errores de medida (*e*) no están correlacionados con la puntuación verdadera (*T*) y (d) el valor esperado de las puntuaciones observadas (*X*) es la puntuación verdadera (*T*). Por su parte, Lord (1980) señala que la mayor parte de estos supuestos se derivan tautológicamente del modelo expresado en la ecuación (2.1).

Idealmente, en TCT deben existir como mínimo dos formas paralelas de un mismo test para comprobar el modelo. Dos formas de un tests son consideradas paralelas si la varianza (σ^2) de los errores (*e*) es la misma para los dos (*j, k*) [$\sigma^2(e_j) = \sigma^2(e_k)$] y si las puntuaciones verdaderas (*T*) obtenidas tras la aplicación de las dos formas (*j, k*) también es igual [$V_j = V_k$].

Finalmente, es conveniente recordar que la TCT se formuló inicialmente como un modelo para explicar la fiabilidad del instrumento de medida (Spear-

man, 1904b, 1907, 1910, 1913), aplicándose más tarde los mismos principios en el estudio de la validez (Gulliksen, 1987).

2.1.1 Teoría Clásica de los Tests y fiabilidad

Cada vez que se aplica un instrumento de medida, se espera que las puntuaciones ofrecidas por el mismo contengan el menor grado posible de errores de medición. Como se indicó previamente, la fiabilidad se ocupa de cuantificar el error de medida no sometido a control, el cual es inevitable en todo proceso de medición. Formalmente, el ***coeficiente de fiabilidad*** permite cuantificar este último. Se define el mismo como la proporción de la variación de *X* (puntuación observada) explicada por *T* (puntuación verdadera), según se denota en la ecuación (2.2).

$$\rho_{xx'} = \frac{\sigma_T^2}{\sigma_X^2} \tag{2.2}$$

En otras palabras, una correlación (ρ) igual a 1 entre dos formas paralelas de un mismo test (x y x') implica estar frente a un instrumento completamente fiable. A medida que $\rho_{xx'}$ se aleja de 1, ello es indicador de estar frente a errores de medición. Tal diferencia permite hacer inferencias sobre la magnitud en que los errores de medición afectan la fiabilidad del instrumento.

Aunque existen distintas formas de evaluar la fiabilidad del instrumento de medida, los métodos *test-retest* y el de *consistencia interna* suelen ser los más utilizados (Muñiz, 1996, 2003). El método *test-retest* consiste en la aplicación de un mismo instrumento a una misma muestra de sujetos en al menos dos momentos diferentes. No existe un criterio único respecto de cual debe ser el lapso adecuado entre la primera y segunda aplicación. Este se establece en función de los fines del instrumento o la investigación (Muñiz, 1996).

Por su parte, el *método de consistencia interna* permite aplicar el instrumento tan sólo una vez. Una forma de evaluar la consistencia interna de un instrumento es mediante el *procedimiento de dos mitades*. Éste consiste, una vez obtenidas las puntuaciones de los sujetos, en la división del test en dos partes equivalentes, las cuales son correlacionadas entre sí. Sin embargo, el procedimiento más utilizado (Muñiz, 1996, 2003) para evaluar la consistencia interna de un test es el *coeficiente Alfa* (α) (Cronbach, 1951), el cual se expresa formalmente según la ecuación (2.3).

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} \sigma_j^2}{\sigma_x^2}\right) \tag{2.3}$$

donde n denota el número de ítems; $\Sigma\sigma_j^2$ la suma de las varianzas de los n ítems y σ_x^2 la varianza de las puntuaciones del test (Muñiz, 2003, p. 54).

2.1.2 Teoría Clásica de los Tests y validez

En cuanto a la validez del instrumento de medida, la TCT propone relacionar la puntuación observada en el test con variables criterio externas al mismo o, lo que es lo mismo, recoger *evidencias basadas en la relación con otras variables*, según se definió en la introducción del apartado 2. Es decir, lo que se analiza es la relación entre la *puntuación global obtenida en la escala* (*T*), expresada habitualmente como la suma de puntuaciones parciales o una media de las mismas, con *variables externas* que guardan directa relación con el constructo evaluado por el instrumento.

2.2. Análisis Factorial de los Ítems

Al mismo tiempo que establece los fundamentos del Modelo Lineal o Teoría Clásica de los Tests (TCT) (Spearman, 1904b), Spearman publica un artículo seminal que sienta las bases de lo que hoy se entiende por modelo de *Análisis Factorial* (AF) (Spearman, 1904a). Cerca de treinta años después, Thurstone (1931, 1947) publica sus trabajos sobre AF, lo cual ayuda a popularizar notablemente este modelo (Bartholomew, 1995). Poco tiempo después, Lawley (1940, 1943a; Lawley y Maxwell, 1971) formula el estimador *máxima verosimilitud*, siendo ésta la primera técnica estadística orientada a la contrastación de hipótesis respecto de la estructura de covarianzas observada, tras la aplicación de un AF (Krane y Slaney, 2005). "En la segunda mitad de la década de 1960, teoría y práctica se encuentran cara a cara. Los analistas factoriales dejan de utilizar este método de manera estrictamente mecánica y comienzan a tratarlo como un medio para poner a prueba hipótesis sobre las covarianzas" (Krane y Slaney, 2005, p. 127). Así, el modelo AF permite contrastar la bondad de ajuste del modelo propuesto, lo cual no era posible desde la perspectiva de la TCT (Bentler, 1980; Bollen, 1989; Bollen y Long, 1993; Jöreskog y Söborn, 1979; Krane y Slaney, 2005; McDonald, 1985, 1999).

En términos generales, el modelo AF contempla una serie de técnicas estadísticas de análisis multivariante diseñadas para investigar la estructura subyacente o dimensiones básicas de un conjunto amplio de variables (Walsh y Betz, 1995) y/o para representar un número amplio de variables observadas en un conjunto menor de constructos o variables (Hair, Anderson, Tatham y Black, 1999), denominados *factores o rasgos latentes*. Éstos últimos son inferidos a partir de la puntuación observada o empírica (Y) obtenida por el sujeto tras contestar uno o varios ítems que dan lugar a la escala utilizada para evaluar un construc-

to psicológico (η) en particular (McDonald, 1999; Muñiz, 2003). Aunque en todos los casos se aplica el mismo modelo general de AF (McDonald, 1999), cuando éste se centra exclusivamente en el análisis de los ítems que componen una escala (p. ej., análisis de la capacidad discriminativa, dificultad de cada ítem, etc.), suele reservarse el nombre de *Análisis Factorial de los Ítems* (AFI) (Ferrando, 1996; McDonald, 1999; Santisteban-Requena, 1990). De aquí en adelante se utilizan indistintamente las siglas AF o AFI para referirse al modelo de *Análisis Factorial de los Ítems*, de manera tal de poder distinguirlo del procedimiento de *Análisis de Componentes Principales* (ACP)[4].

En la actualidad se distinguen dos formas de AF: *Análisis Factorial Exploratorio* (AFE) y *Análisis Factorial Confirmatorio* (AFC). A nivel descriptivo, la principal diferencias entre ambos tipos de AF radica en el hecho de que AFE se utiliza para *detectar* fuentes latentes de variación y covariación entre las variables observadas, mientras que AFC se aplica cuando el investigador cuenta con una *hipótesis previa acerca de la estructura* de las variables latentes y quiere comprobar la misma (Martínez-Arias, 1996). La Tabla 1 resume lo que Bollen (1989) identifica como características principales de AFE y AFC.

En síntesis y de acuerdo a la Tabla 1, mientras que AFE se utiliza para examinar libremente las fuentes latentes de variación y covariación entre variables observadas, AFC requiere de un modelo inicial detallado e identificado que debe ser contrastado (Bollen, 1989). "En áreas sustantivas donde aun se conoce

[4] Aunque ambos procedimientos se utilizan para reducir una matriz de datos en su mínima expresión, el objetivo principal de todo *Análisis de Componentes Principales* (ACP) consiste en explicar la varianza total de las variables resultantes (componentes), en tanto que el objetivo fundamental del *Análisis Factorial* (AF) consiste en reproducir la matriz de correlaciones teórica a partir de los datos, en función de las comunalidades observadas (factores). Es decir, mientras que ACP busca reducir un conjunto de datos para establecer un modelo matemático hipotético, AF parte ya de un modelo y utiliza los datos para reproducir el mismo (Hair *y col.* 1999; Krane y Slaney, 2005; Martinez-Arias, 1996).

poco, el análisis factorial exploratorio puede ser muy valioso ya que permite sugerir patrones subyacentes en los datos. Sin embargo, si existen hipótesis plausibles sobre la estructura de un modelo, entonces el análisis factorial exploratorio puede frustrar las tentativas para probar tales ideas" (Bollen, 1989, p. 228). En este último caso, es preferible utilizar un AFC.

Tabla 1: Principales características y diferencias entre Análisis Factorial Exploratorio (AFE) y Análisis Factorial Confirmatorio (AFC).

Análisis Factorial Exploratorio (AFE)	*Análisis Factorial Confirmatorio (AFC)*
• No se especifica un modelo previo que relacione las variables latentes y observadas. • El número de variables latentes no se fija antes del análisis. • Por regla general, todas las variables latentes influyen sobre todas las variables observadas. • El efecto de las variables latentes sobre las variables observadas es libre; es decir, no es especificado por el analista. • Los errores de medida (δ) no están correlacionados entre sí. • La covariación (COV) de las variables latentes no se especifica previamente por el analista. • No es necesario identificar los parámetros del modelo.	• Se construye un modelo previo que especifica la(s) relación(es) entre las variables latentes y observadas. • El número de variables latentes se especifica por el analista antes de realizar el análisis. • La influencia de las variables latentes sobre las variables observadas está fijada por el analista. • El efecto de las variables latentes sobre las variables observadas puede ser fijado a 0 o a otro valor constante (p. ej., 1) • Los errores de medida (δ) pueden estar correlacionados. • La covariación (COV) de las variables latentes puede ser estimado o fijado en determinado valor. • Se requiere la identificación de los parámetros del modelo.

Adaptado de Bollen (1989).

Formalmente, cuando se asume que únicamente es necesario un factor (η), el modelo general de AFI se expresa a través de la ecuación (2.4)

$$Y_{ij} = \mu_i + \lambda_i \eta_j + e_{ij} \qquad (2.4)$$

donde Y_{ij} representa la puntuación observada en el ítem *i* para el sujeto *j*, η_j es la puntación del sujeto *j* en el factor o rasgo latente, μ_i es el intercepto que cambia de ítem a ítem, λ_i es la pendiente de regresión del rasgo latente o factor (η) sobre el ítem *i* observado y e_{ij} es el término de error que contiene tanto los errores de especificación como los de medida.

De acuerdo a lo expresado en la ecuación (2.4), en este modelo se asume que, dentro de la población evaluada: (a) el valor esperado de los errores es cero, (b) los errores no están correlacionados con el rasgo latente $\left[corr(\eta, e_i) = 0\right]$ y (c) que los errores no están correlacionados entre sí $[corr(e_i, e_{i'}) = 0]$. Asimismo, en este modelo, la media y la varianza de factor no están identificados (no pueden ser estimados) y por tanto se fijan por convención a cero y uno, respectivamente.

En resumen, los parámetros del modelo expresado en la ecuación (2.4) son, para cada ítem, μ_i, λ_i, y ψ_i^2, donde este último parámetro es la varianza de los errores para el ítem *i*.

2.2.1 Análisis Factorial de lo Ítems y Fiabilidad

Mientras que desde el modelo de TCT se puede utilizar el coeficiente alfa de Cronbach [ver ecuación (2.3)] para evaluar la consistencia interna de la prueba, dentro del modelo AFI ésta puede ser estimada calculando el coeficiente omega (ω), según se expresa en la ecuación (2.5) (McDonald, 1999).

$$\omega = \frac{\left(\sum_{i=1}^{p} \lambda_i\right)^2}{\left(\sum_{i=1}^{p} \lambda_i\right)^2 + \sum_{i=1}^{p} \psi_i^2} \tag{2.5}$$

donde λ_i es la pendiente de regresión del rasgo latente o factor (η) sobre el ítem *i* observado y ψ_i^2 es la varianza de los errores del ítem *i*. En AFI, λ_i y ψ_i^2 se denominan generalmente como *carga factorial* y *unicidad*, respectivamente.

De acuerdo a Muñiz (2003), para los mismos datos, el valor de ω suele ser superior al de α. En todo caso, frente a ítems paralelos, ambos coeficientes deben ser iguales entre sí e iguales al coeficiente de fiabilidad del test ($\rho_{xx'}$).

2.2.2 Análisis Factorial de lo Ítems y validez

Al igual que en la TCT, dentro del modelo AFI también se pueden examinar *evidencias basadas en la relación con otras variables*. De todos modos, en este caso se estiman las correlaciones entre el factor o rasgo latente (η) y los criterios externos relevantes –no entre la puntuación total del test y las variables externas relevantes, como ocurre en el modelo TCT. Otra diferencia entre los modelos TCT y AFI es el hecho que dentro de éste último es posible evaluar la *validez estructural* del constructo psicológico medido por el test, gracias a la valoración de la *bondad de ajuste* del modelo factorial propuesto por el investigador. Es decir, se puede estimar la *evidencia basada en la estructura interna* de la prueba. Por tanto, dentro del modelo AFI, la presente investigación se centra en el análisis del efecto del número de opciones de respuesta sobre la *validez estructural* (evidencia basada en la estructura interna) y *la relación con otras variables* (evidencia ba-

sada en la relación con otras variables). Lo anterior se lleva a cabo, de acuerdo a los criterios y procedimientos que se exponen en el capítulo 4.

2.3. Teoría de Respuesta a los Ítems

Los orígenes de la TRI se remontan a los trabajos pioneros de Lawley (1943b, 1944), Lazarsfeld (1959), Richardson (1936) o Tucker (1946) (Hambleton y Swaminathan, 1985; Lord, 1980; Maydeu-Olivares y McArdle, 2005). Por su parte, Birnbaum, Lord y Rasch ayudan a formalizar la TRI. Entre otras aportaciones, Birnbaum (1957, 1968) introduce los fundamentos del modelo *logístico*, Lord (1952, 1953a, 1953b, 1953c) propone el modelo de *ojiva normal*, y Rasch (1960) introduce un caso especial del modelo logístico, con atractivas propiedades matemáticas, el cual se identifica con su propio nombre. De todos modos, los libros *Statistical Theories of Mental Tests Scores* (Lord y Novick, 1968) primero, y sobre todo *Applications of Item Response Theory to Practical Testing Problems* (Lord, 1980) después, son las síntesis por excelencia de los fundamentos de la TRI (Muñiz, 1997, 2003). Por otro lado, la década de 1980 está marcada por la difusión de los modelos de TRI, no solo por la publicación de libros como el de Lord (1980), sino también por la expansión en el uso de los ordenadores personales y la consiguiente aparición de paquetes estadísticos capaces de estimar modelos de TRI (Hambleton y Swaminathan, 1985; Muñiz, 1997; Van der Linden y Hambleton, 1997). Naturalmente, son muchos otros los autores que han aportado al desarrollo de la TRI. Aquí se han mencionado solo algunos de los más relevantes, no con otro fin que contextualizar históricamente el tema tratado en este apartado.

A continuación se ofrece una breve introducción respecto de los fundamentos del modelo TRI, basada principalmente en Hambleton y Swaminathan

(1985), Lord (1980) y Martínez-Arias (1996). Para aquellos que deseen profundizar este tema, también pueden consultar Maydeu-Olivares y McArdle (2005), McDonald (1999), Muñiz (1997) o Van der Linden y Hambleton (1997).

De acuerdo a lo tratado en los apartados 2.1 y 2.2, los modelos de *Teoría Clásica de los Test* (TCT) y *Análisis Factorial de los Ítems* (AFI) comparten como denominador común el hecho de estar basados en un *modelo lineal* (Gulliksen, 1987; Lord y Novick, 1968; Spearman, 1904a, 1904b, 1907, 1913). A pesar de su gran difusión, ambos modelos presentan dos problemas fundamentales: (a) sus mediciones dependen del instrumento de medida utilizado y (b) las propiedades de estos últimos son dependientes de los sujetos evaluados (Lord, 1980; Martínez-Arias, 1996; Muñiz, 1996, 1997, 2003). La *Teoría de Respuesta a los Ítems* (TRI) procura resolver ambos problemas, estableciendo modelos que intentan asegurar que: (a) las mediciones obtenidas no varíen en función del instrumento utilizado, es decir, que sean invariantes respecto al test, (b) disponer de instrumentos de medida cuyas propiedades no dependan de los objetos medidos o, lo que es lo mismo, sean independientes de los sujetos evaluados (invariantes respecto de los sujetos) y (c) proporcionar un modelo de respuesta (Hambleton y Swaminathan, 1985; Lord, 1980; Martínez-Arias, 1996; Muñiz, 1997). Adicionalmente, el modelo TRI tiene en cuenta la naturaleza categórica de los ítems y por tanto modeliza la probabilidad de obtener cada patrón de respuestas. En particular, el modelo TRI se apoya en dos supuestos fundamentales. Por un lado, se asume que la probabilidad de responder a los ítems depende de solo una variable latente (supuesto de *unidimensionalidad*), aunque también puede estar influido por más de una. Por otro lado, se asume que las probabilidades de responder a los ítems son independientes, para un nivel fijo del rasgo latente (supuesto de *independencia local*).

Existen distintos modelos posibles dentro de la TRI. Estos pueden estar basados en algún tipo de función logística o función de ojiva normal con distinto número de parámetros para cada ítem (Lord, 1980; Martínez-Arias, 1996; Muñiz, 1996, 1997). Dado que en este trabajo se utilizan preferentemente escalas tipo Likert, es conveniente utilizar un modelo que tenga en cuenta la naturaleza ordinal de los datos. Maydeu-Olivares (1991, 2005a, 2005b) y Ferrando (1996) señalan que el modelo más adecuado para este tipo de datos es el *modelo gradual de Samejima* o *modelo de respuesta graduada de Samejima* (Samejima, 1969). En dicho modelo, la probabilidad de obtener la categoría k para un nivel fijo en el rasgo latente η viene dada por la ecuación (2.6)

$$\Pr\left(Y_i = k|\eta\right) = \begin{cases} 1 - G\left(\alpha_{i,1} + \beta_i\eta\right) & \text{if} \quad k = 0 \\ G\left(\alpha_{i,k} + \beta_i\eta\right) - G\left(\alpha_{i,k+1} + \beta_i\eta\right) & \text{if} \quad 0 < k < K-1 \\ G\left(\alpha_{i,m-1} + \beta_i\eta\right) & \text{if} \quad k = K-1 \end{cases} \tag{2.6}$$

donde $G\left(\alpha_{i,k} + \beta_i\eta\right)$ es igual a una distribución normal estándar

$$\Phi\left(\alpha_{i,m-1} + \beta_i\eta\right) = \int_{-\infty}^{\alpha_{i,m-1}+\beta_i\eta} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \tag{2.7}$$

o a una distribución logística estándar

$$\Psi\left(\alpha_{i,m-1} + \beta_i\eta\right) = \frac{1}{1 + \exp\left[-\left(\alpha_{i,m-1} + \beta_i\eta\right)\right]} \tag{2.8}$$

Así, dentro del modelo de respuesta graduada de Samejima, para cada ítem hay un parámetro de pendiente β_i y m - 1 parámetros de intercepto $\alpha_{i,k}$.

Por otro lado, cuando se utilizan funciones normales, es posible relacionar el modelo de respuesta graduada de Samejima con el modelo AFI. Antes de

describir dicha relación, conviene recordar algunas diferencias generales entre los modelos AFI y TRI.

Dentro del modelo AFI (ver apartado 2.2) no es necesario asumir ningún tipo de distribución específico para los rasgos latentes (Maydeu-Olivares, 2005b; Maydeu-Olivares, Gallardo y Kramp, 2004). Asimismo, en AFI se modelizan solo las matrices de medias y covarianzas de los ítems observados, en tanto que los momentos de orden superior de los datos no son modelizados. Tal y como señala Ferrando (1996): "…la principal motivación que ha guiado tradicionalmente a los usuarios del AF ha sido la de determinar la estructura de covarianzas entre variables (y bautizar factores) pero muy pocos investigadores se han interesado por la estimación de las puntuaciones factoriales" (Ferrando, 1996, p. 570). Por el contrario, la principal motivación de quienes recurren al modelo TRI consiste en medir el/los rasgo/s latente/s (θ)[5] de interés, con objeto de determinar el nivel presente en el sujeto de dicho/s rasgo/s (θ). Asimismo, se modelizan todos los momentos conjuntos de los datos, por medio de algún tipo de función (p. ej., logística o de ojiva normal).

En cuanto a las similitudes, la relación entre AFI y TRI surge desde la ecuación (2.4), donde se formalizó el modelo general de AFI como $Y_{ij} = \mu_i + \lambda_i \eta_j + e_{ij}$. En AFI se asume que Y_{ij} es una variable observada *continua*. En cambio, dentro del marco de TRI y específicamente desde el modelo de respuesta graduada de Samejima (Samejima, 1969), se asume que las variables Y_{ij} no son observadas, sino únicamente sus categorizaciones Z_{ij}. Formalmente, si Z_i es un ítem con K categorías $0, 1, \ldots, k, \ldots, K-1$, el modelo de respuesta gra-

[5] Por razones de forma y distinguir claramente entre modelos, hemos reservado el concepto de *factor* (η) para referirnos a los constructos evaluados desde el modelo AFI, mientras que el concepto de *rasgo latente* (θ) para dar cuenta de los constructos evaluados bajo los supuestos del modelo TRI.

duada de Samejima con ojivas normales puede obtenerse a partir del modelo AFI (ver ecuación (2.4)) mediante la relación

$$Z_i = \begin{cases} 0 & \text{si} \quad Y_i \le \tau_{i,1} \\ k & \text{si} \quad \tau_{i,k} < Y_i \le \tau_{i,k+1} \\ K-1 & \text{si} \quad Y_i > \tau_{i,K} \end{cases} \tag{2.9}$$

Así, de acuerdo a (2.6), el modelo especifica un conjunto de *K-1* umbrales (τ), los cuales cambian de ítem a ítem. Asimismo, el modelo asume que los errores aleatorios e_{ij} y el rasgo latente η se distribuyen normalmente. Dado que no es posible estimar todos los parámetros que aparecen en este modelo, es necesario fijar los interceptos μ_i a cero y fijar $\psi_i^2 = \sqrt{1-\lambda_i^2}$ para identificarlo.

En resumen, el modelo de respuesta graduada de Samejima, cuando se utilizan ojivas normales, puede ser descrito utilizando *K*-1 umbrales $\tau_{i,k}$ y una carga factorial λ_i por ítem, o equivalentemente *K*-1 interceptos $\alpha_{i,k}$ y una pendiente β_i por ítem. La relación entre las dos parametrizaciones del modelo está dada por $\alpha_{i,k} = \frac{-\tau_{i,k}}{\sqrt{1-\lambda_i^2}}, \beta_i = \frac{\lambda_i}{\sqrt{1-\lambda_i^2}}$ (McDonald, 1997).

2.3.1 Fiabilidad y Validez desde el modelo de Teoría de Respuesta a los Ítems

Desde el modelo de la TRI, es habitual evaluar la precisión (fiabilidad) de los constructos psicológicos evaluados por medio del análisis de la *función de información* aportada por el instrumento de medida (Samejima, 1969; Lord, 1980). Ésta se caracteriza por ser una función *no lineal* de los parámetros de los ítems.

Así, a diferencia de los modelos TCT y AFI, la precisión del instrumento de medida no es constante para todos los niveles del rasgo latente estudiado (θ). En otras palabras, mientras que para TCT y AFI la fiabilidad se corresponde con un número, en TRI ésta es una función del nivel del rasgo latente (θ).

En el entendido de que este trabajo de Tesis de Doctorado tiene por objeto comparar el efecto del número de opciones de respuesta sobre las propiedades psicométricas de las pruebas de personalidad, desde la óptica de los modelos TCT, AFI y TRI, se ha estimado conveniente utilizar como medida de precisión para TRI el coeficiente omega (ω) descrito anteriormente para AFI. Esto es posible gracias a la relación entre el modelo AFI cortado con umbrales y el modelo de respuesta graduada de Samejima. Pensamos que de esta manera se cuenta con un medio que permite comparar de forma directa la fiabilidad (precisión) del instrumento de medida desde los tres modelos psicométricos utilizados como referentes (TCT, AFI y TRI). Para el caso particular del modelo TRI, el coeficiente omega (ω) queda definido como un promedio de la función de información expresada por el test (McDonald, 1999).

Por lo tanto, la fiabilidad (precisión) y la validez de los instrumentos de medición pueden ser evaluadas en TRI de acuerdo a los mismos criterios expuestos en los apartados 2.1 y 2.2. La diferencia entre los modelos expuestos para AFI y TRI, radica en el hecho de que para este último los datos son tratados categóricamente y no de forma continua. Asimismo y como consecuencia del tratamiento categórico de los datos, el ajuste estructural del modelo se estima a partir de la bondad de ajuste de la matriz de correlaciones policóricas del modelo subyacente, según se indica en el capítulo 4.

2.4. Modelos de Ecuaciones Estructurales

Los *Modelos de Ecuaciones Estructurales* (SEM), conocidos también como *Modelos de Estructuras de Covarianzas, Path Analysis con Variables Latentes* o *Ecuaciones Estructurales con Variables Latentes* (Bollen, 1989; Gómez, 1996; Martínez-Arias, 1996; McDonald, 1999), utilizan estructuras de covarianzas (o correlaciones) para estimar y verificar modelos estadísticos, sustentados en la teoría. En términos generales, SEM puede definirse como una *metodología* estadística orientada al análisis de la dependencia-independencia de un conjunto de variables medidas y/o dirigido al análisis de los factores comunes que dan forma a subconjuntos de variables (Bollen, 1989; McDonald, 1999). "Los modelos de estructuras de covariancia constituyen una metodología estadística de tipo confirmatorio, en la que a partir de la teoría se formulan explicaciones causales sobre variables latentes y a partir de los datos se evalúa la consistencia de las relaciones hipotetizadas" (Gómez, 1996, p. 463). O, en palabras de Bollen (1989), "podemos mirar estos modelos de diversos modos. Son ecuaciones de regresión con supuestos menos restrictivos, que permiten errores de medida tanto en las variables criterio (independientes) como en las variables dependientes. Consisten en análisis factoriales que permiten efectos directos e indirectos entre los factores. Habitualmente incluyen múltiples indicadores y variables latentes. Resumiendo, engloban y extienden los procedimientos de regresión, el análisis econométrico y el análisis factorial" (Bollen, 1989, p. V)

A nivel descriptivo, SEM es un proceso compuesto por una serie de etapas orientadas a minimizar la diferencia entre las covarianzas muestrales y las covarianzas predichas por el modelo propuesto. SEM busca modelizar la matriz de varianzas – covarianzas de las variables observadas. Para ello, asume que la matriz de covarianzas poblacional de las variables observadas depende de un

vector de parámetros a estimar (Bollen, 1989). Formalmente, lo anterior se expresa mediante la ecuación (2.10).

$$\Sigma = \Sigma(\theta) \tag{2.10}$$

donde Σ denota la matriz de covarianzas poblacional de variables observadas, θ es un vector que contiene los parámetros del modelo y $\Sigma(\theta)$ es la matriz de covarianzas escrita como una función de θ.

En el entendido de que los parámetros del modelo se desconocen, se utiliza una matriz de covarianzas muestral (S) como estimación no sesgada de Σ y se estima el vector θ. Esto último se consigue minimizando alguna función de discrepancia F[S, $\Sigma(\theta)$], a partir de la cual se establecen índices de ajuste que permiten evaluar la bondad de ajuste del modelo evaluado. "La hipótesis fundamental para estos procedimientos de ecuaciones estructurales es que la matriz de covarianzas de las variables observadas es una función de un conjunto de parámetros. Si el modelo es correcto y si se conocen los parámetros, es posible reproducir la matriz de covarianzas poblacional de forma exacta" (Bollen, 1989, p. 1).

Antes de precisar con más detalle las etapas que dan cuenta de un proceso SEM, conviene recordar brevemente sus orígenes[6] y algunos de los elementos que permiten expresar formalmente un modelo a contrastar. Respecto de lo primero, Bentler (1980) y Bollen (1989) señalan que SEM surge de la confluencia de tres líneas de investigación: psicometría, econometría y biometría. Desde la *psicometría* se rescatan los conceptos de variable latente y de error de medida, asociados al estudio del Análisis Factorial y la Teoría de la Fiabilidad. Por su parte, desde la *econometría* se hace uso del análisis de las influencias direcciona-

[6] Para aquellos que deseen profundizar en este tema, pueden consultar Bentler (1986), Bollen (1989) o Goldberger (1972).

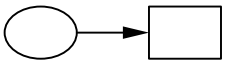
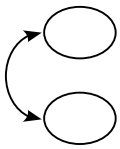
les y simultáneas de un conjunto de variables, mientras que desde la *biometría* se toma el desarrollo de esquemas sobre representación y estimación, lo cual da lugar a los análisis de sendero o Path-Analysis. El primer modelo práctico de SEM se conoce como modelo *JKW* y fue desarrollado por Jöreskog (1973), Keesling (1972) y Wiley (1973). Por su parte, Jöreskog y Söborn crean durante la década de 1970 el programa LISREL, gracias al cual se extiende rápidamente el uso de SEM. En la actualidad, éste va en su versión octava (Jöreskog y Sörbom, 2005). Otros programas destacados son AMOS (SPSS, 2005), EQS (Bentler, 1995), Mplus (L.K. Muthén y Muthén, 2005), etc. Por nuestra parte, hemos decidido trabajar con Mplus, principalmente por su plasticidad y la sencillez de su interfaz.

En cuanto a los elementos que permiten expresar formalmente un modelo SEM, estos se refieren al tipo de variables implicadas, la relación que existe entre estas y los parámetros estructurales que dan cuenta del modelo a contrastar. Respecto de las ***variables***, estas pueden ser observadas, latentes, exógenas y/o endógenas. Se identifican como *observadas* si pueden ser medidas directamente y como *latentes* si se infieren a partir de las primeras. Por su parte, una variable se considera *endógena* si recibe la influencia direccional de otra variable presente en el modelo, mientras que se clasifica como *exógena* si no recibe influencia por parte de ninguna variable.

En cuanto al tipo de ***relación*** que se establece entre las variables, esta puede ser direccional o no direccional. Si la relación se define como *direccional*, da lugar a un coeficiente de regresión lineal, en tanto que si la relación se especifica como *no direccional*, da cuenta de valores de covarianza entre las variables. Ambos valores (direccional y no direccional) constituyen los ***parámetros del modelo***, cuya estimación es el primer objetivo del análisis SEM. Finalmente, es necesario establecer también el valor de los parámetros, el cual puede ser fijo o

libre. Un parámetro se considera *fijo* cuando se especifica de antemano su valor, el cual suele ser 0, indicando que no hay relación entre las variables (o algún grado distinto de 0, dependiendo del valor especificado). Un parámetro se considera *libre* cuando su valor es desconocido y se estima a partir del análisis de los datos. En términos gráficos, lo anterior puede ser representado por un *Diagrama de Sendero* o *Path Diagram*, en función de los símbolos que se resumen en la tabla 2.

Tabla 2: Símbolos básicos para realizar un *Path Diagram*.

	Un cuadrado o rectángulo denota una variable observada o manifiesta.
	Un círculo o elipse indica una variable no observada o latente.
	Se utilizan flechas para indicar la relación 'causal' entre variables. Se asume que la variable en la base de la flecha causa la variable que está en su punta.
	Una línea curva con flechas que conectan simultáneamente dos variables indica covariación entre las mismas.

2.4.1 Etapas de un Modelos de Ecuaciones Estructurales

Basado principalmente en Bollen (1989) y Gómez (1996), a continuación se ofrece una descripción general de las etapas que dan lugar a un proceso de análisis mediante SEM. Para una revisión detallada de los tópicos que siguen, se

puede consultar Bentler (1980, 1986), Bollen (1989; Bollen y Long, 1993), Goldberger (1972) o Gómez (1996).

En esencia, SEM es un proceso de análisis estadístico que consta de cinco pasos: (a) especificación del modelo, (b) identificación, (c) estimación, (d) evaluación del ajuste del modelo y (e) reespecificación (Bollen, 1989; Bollen y Long, 1993). "Los procedimientos de análisis de estructuras de covariancias tratan de explicar las relaciones entre las variables observables mediante la formulación de modelos que sean sustantivamente explicativos y al mismo tiempo parsimoniosos con respecto a la estructura de los datos" (Gómez, 1996, p. 464). El análisis SEM comienza con la ***especificación formal de un modelo*** a contrastar, el cual está basado en la teoría. En este se establecen las variables que lo componen, las cuales pueden ser observables, latentes, exógenas y endógenas. Al mismo tiempo, se especifica el tipo de relación que existe entre éstas. Como se indicó anteriormente, dicha relación puede ser direccional o no direccional y da cuenta de los parámetros del modelo. Tales parámetros pueden ser definidos como fijos, libres o restringidos.

La fase de ***identificación*** del modelo consiste en el análisis de la correspondencia entre la información que debe obtenerse (parámetros libres) y la información de que se dispone (varianzas-covarianzas o correlaciones de las variables observadas). "Para que un parámetro sea identificado es preciso que pueda expresarse como una función de las varianzas y covarianzas muestrales y que esta función conlleve una solución única para el valor del parámetro" (Gómez, 1996, p. 466). Una vez identificado o sobreidentificado el modelo es posible entrar en la tercera fase de ***estimación de los parámetros***, donde se estiman los parámetros libres mediante métodos iterativos capaces de generar una matriz de varianzas-covarianzas Σ lo más parecida posible a la matriz de varianzas-covarianzas obtenida (S) a partir de los datos utilizados. Entre otros,

los métodos iterativos más utilizados suelen ser *máxima verosimilitud* (ML), *mínimos cuadrados no ponderados* (ULS) o *mínimos cuadrados generalizados* (GLS) (Bollen, 1989; Bollen y Long, 1993; L. K. Muthén y Muthén, 2001, 2004).

La fase de ***evaluación del ajuste del modelo*** consiste en comprobar el grado en que coinciden las matrices S y Σ, para lo cual suele utilizarse como medida de **ajuste global** el estadístico chi-cuadrado (X^2), tanto bajo supuestos de normalidad como sin estos supuestos (Bollen, 1989; Bollen y Long, 1993; Muthén, 1993; Satorra y Bentler, 1994; Muthén, du Toit y Spisic, 1997; L. K. Muthén y Muthén, 2001, 2004) . "El estadístico X^2 expresa el grado en que un modelo especificado ajusta a los datos comparado con el modelo saturado y cumple la función de evaluar en qué medida el modelo representa las relaciones observadas" (Gómez, 1996, p. 468). Como es bien sabido, el índice X^2 es la única medida de bondad de ajuste en SEM fundamentada estadísticamente (Bollen, 1989; Bollen y Long, 1993; Hair *y col.*, 1999; Marsh, Hau y Grayson, 2005). Así, "…valores de chi-cuadrado bajos, que resultan en niveles de significación mayores que 0.05 ó 0.01, indican que las matrices de entrada previstas y efectivas no son estadísticamente diferentes…Si embargo, incluso la no significación estadística no garantiza haber identificado el modelo <<correcto>>, sino sólo que este modelo propuesto se ajusta bien a las varianzas y covarianzas observadas" (Hair *y col.*, 1999, p. 680).

En el entendido que a menudo los modelos SEM muestran un pobre nivel de ajuste absoluto (X^2), es común encontrar en la literatura SEM la recomendación de evaluar la **bondad de ajuste aproximada** del modelo (Bollen, 1989; Bollen y Long, 1993; Hair *y col.*, 1999; Marsh *y col.*, 2005). De acuerdo a Hu y Bentler (1995), se han formulado distintos índices para evaluar la bondad de ajuste aproximada de un modelo SEM, los cuales pueden clasificarse en (a) índices de ajuste absoluto o (b) índices de ajuste incremental. Los ***índices de ajuste***

absoluto se refieren al grado en que un modelo hipotetizado es capaz de reproducir los datos muestrales; entre otros, cabe mencionar el *parámetro de no centralidad* (NCP), el *índice de bondad de ajuste* (GFI), el *índice de residuo cuadrático medio* (RMSR), el *índice de error de aproximación cuadrático medio* (RMSEA), etc. (Bollen, 1989; Bollen y Long, 1993; Hair *y col.*, 1999; Marsh *y col.*, 2005). Por su parte, los ***índices de ajuste incremental o comparativo*** evalúan el aumento proporcional en ajuste comparando el modelo hipotético con otro que presenta restricciones, generalmente el modelo nulo en el que todas las variables observadas están relacionadas entre si. Entre otros, está el *índice ajustado de bondad de ajuste* (AGFI), *índice de ajuste normal* (NFI), *índice Tucker-Lewis* (TLI), *índice de ajuste comparado* (CFI), etc. (Bollen, 1989; Bollen y Long, 1993; Hair *y col.*, 1999; Marsh *y col.*, 2005).

En lo que a este trabajo de Tesis de Doctorado toca, los estudios realizados se evalúan por medio del índice de bondad de ajuste global X^2, sin supuestos de normalidad (según se indica en el capítulo 4, subapartado 4.4), y los índices de bondad de ajuste aproximada RMSEA (Steiger, 1990), TLI (Tucker y Lewis, 1973) y el CFI (Bentler, 1990).

Respecto de los valores de referencias para valorar los índices de bondad de ajuste aproximada antes indicados, para ***RMSEA*** se considera *adecuado* un valor comprendido entre 0 y 0.05, mientras que *aceptable* un valor que oscile entre 0.05 y 0.08 (Browne y Cudeck, 1993). Por su parte, tanto para el índice ***TLI*** como para el índice ***CFI*** se consideran *adecuados* los valores que están por sobre un umbral de 0.9, siendo 1.00 reflejo de un ajuste perfecto (Bentler, 1990; Hair y *col.*, 1999; Tucker y Lewis, 1973).

2.4.2 Estimación del modelo graduado de Samejima desde un marco de Modelos de Ecuaciones Estructurales

Como se señaló en el apartado 2.3, el modelo TRI de respuesta graduada de Samejima puede obtenerse como una extensión del modelo AFI expuesto en la ecuación (2.4), mediante la ecuación (2.6). La tabla 3 recuerda ambas ecuaciones.

Tabla 3: Relación entre los modelos de Análisis Factorial de los Ítems (AFI) y Teoría de Respuesta a los Ítems (TRI).

$Y_{ij} = \mu_i + \lambda_i \eta_j + e_{ij}$	Ecuación (2.4)
$Z_i = \begin{cases} 0 & \text{si} \quad Y_i \le \tau_{i,1} \\ k & \text{si} \quad \tau_{i,k} < Y_i \le \tau_{i,k+1} \\ K-1 & \text{si} \quad Y_i > \tau_{i,K} \end{cases}$	Ecuación (2.6)

Como se señaló en el apartado 2.3 el modelo de respuesta graduada de Samejima se obtiene combinando ambas ecuaciones.

Dentro de un enfoque SEM, el modelo gradual de Samejima con ojivas normales se estima como sigue. Primero, se estiman los umbrales y las correlaciones policóricas que derivan de los datos. Las correlaciones policóricas son correlaciones entre dos variable no observadas Y_i. Luego se estiman los parámetros restantes del modelo a partir de la matriz de correlaciones policóricas obtenida en el paso previo.

Por lo tanto, desde SEM es posible relacionar los tres modelos psicométricos de referencia utilizados en esta investigación (TCT, AFI y TRI), de acuerdo a los siguientes criterios generales. Para TCT, se modelizan las puntuaciones totales de las escalas (X) con las variables externas a partir de la matriz de corre-

laciones producto-momento. Desde el modelo AFI, esta relación se establece entre los ítems, tratados como datos continuos (Y), y las variables externas, también a partir de la matriz de correlaciones producto-momento. Finalmente, para el modelo TRI se utilizan los ítems de las escalas, tratados categóricamente (Z), los cuales se relacionan primero entre sí mediante matrices de correlaciones policóricas y luego con las variables externas, mediante matrices de correlaciones poliseriales.

3. *Estado de la cuestión*

A continuación se ofrece una descripción sobre el *estado de la cuestión*, respecto del estudio del efecto sobre las propiedades psicométricas de los cuestionarios de personalidad, desde una *perspectiva centrada en el sujeto* (Torgerson, 1962).

El primer trabajo que aborda específicamente la relación entre número de opciones de respuesta y las propiedades psicométricas de los instrumentos que evalúan las actitudes o la personalidad corresponde a **Symonds** (1924). Este autor *reflexiona* acerca de cuál cree que debe ser el número óptimo de alternativas de respuesta, capaz de garantizar una adecuada fiabilidad inter-jueces, en este tipo de instrumentos. Concluye que siete es el número óptimo de opciones, por cuanto permite garantizar la obtención de niveles aceptables de fiabilidad, respetando los índices habituales de fiabilidad inter-jueces observado en este tipo de pruebas —en torno a .60 (Symonds, 1924). Este artículo abre el debate sobre el tema que nos ocupa, el cual sigue vigente hasta nuestros días.

Pemberton (1933) y **Champney y Marshall** (1939) continúan las reflexiones de Symonds (1924) acerca del grado de *especificidad* necesario de utilizar al momento de construir una escala de respuesta. Ambos argumentan a favor del uso de escalas continuas graficas, más finas y con un mayor gradiente de respuesta, evaluado en milímetros, que las sugeridas por Symonds (1924) (Champney y Marshall, 1939; Pemberton, 1933)[7]. A pesar de ello, hay que espe-

[7] Tras la aparición del método aditivo de Likert (1932), el uso de escalas gráficas de respuesta pierde fuerza, al menos en que al ámbito de la evaluación de la personalidad se refiere. A pesar de ello, a partir de la década de 1980 estas últimas comienzan a ser analizadas nuevamente, respecto de su efecto sobre las propiedades psicométricas de este tipo de instrumentos de evaluación (p. ej., Cicchetti *y col.*, 1985; Ferrando, 1995, 1999, 2000; Preston y Colman, 2000).

rar hasta la década de 1960 para que este cobre fuerza. **Peabody** (1962) es el primero en responder de forma empírica la propuesta inicial de Symonds (1924). Contrariamente a este último, plantea que basta utilizar solo dos opciones de respuesta para garantizar una adecuada fiabilidad del instrumento de medida. Llega a esta conclusión tras colapsar una escala compuesta inicialmente por 6 alternativas a otra de 2 opciones de respuesta y observar que la fiabilidad del instrumento no varía sustantivamente entre uno y otro tipo de formato de respuesta. Así, sugiere utilizar un número reducido de opciones de respuesta, en tanto ofrece mayor parsimonia como modelo. Complementariamente, estudia también los efectos del gradiente y dirección de la respuesta en escalas tipo Likert. Plantea que cuando se responde a una escala que evalúa actitudes, se genera una puntuación "compuesta" que resulta de combinar la *dirección* (de acuerdo o en desacuerdo) con la *intensidad* de la respuesta (magnitud en que el sujeto está de acuerdo o en desacuerdo). Concluye que un 10% de la puntuación obtenida puede atribuirse a la intensidad de la respuesta, mientras que un 70-80% es resultado de la dirección (Peabody, 1962).

Komorita y Graham (1965) estudian luego el efecto de la homogeneidad de los ítems, el número de opciones de respuesta (2 y 6 alternativas de respuesta) y el número de ítems sobre la fiabilidad de la escala. El grado de homogeneidad de los ítems lo definen a partir del tamaño de las cargas factoriales, sobre un factor único. Indican que el coeficiente alfa es independiente del número de opciones de respuesta, en escalas con ítems homogéneos. A su vez, señalan que el tamaño de la carga factorial juega un rol mediador en la relación entre el diseño de la escala y su fiabilidad. Indican que las escalas con ítems homogéneos se ven poco afectadas por el número de opciones utilizado, no así las escalas con ítems heterogéneos. En este último caso, en función de la variabilidad implícita en los ítems de la escala, al aumentar el número de opciones de respuesta y el número de ítems, aumenta también la fiabilidad de la escala. En to-

do caso, señalan que si se aumenta el número de ítems infinitamente, llega un punto en que el efecto del número de opciones de respuesta queda anulado. Así, fiabilidad y número de categorías de respuesta, en rigor, pueden ser considerados elementos independientes. En la práctica, Komorita y Graham (1965) concluyen que el aumento en fiabilidad para escalas heterogéneas no significa necesariamente una mayor validez de la escala, lo cual implica una paradoja que requiere ser estudiada en mayor profundidad.

Matell y Jacoby (1971) confirman parcialmente las conclusiones de Komorita y Graham (1965) y apoyan lo señalado por Peabody (1962). Estudian escalas con opciones de respuesta que contienen desde 2 a 19 alternativas. Sugieren que basta utilizar 2 ó 3 categorías de respuesta, pues la fiabilidad o la validez no varían sustantivamente entre las distintas condiciones evaluadas. Llegan a esta conclusión tras colapsar las escalas con mayor número de opciones de respuesta en escalas compuestas por 2 ó 3 alternativas y observar que la fiabilidad o la validez de la prueba no disminuyen sustantivamente. A pesar de lo anterior, hay que observar con precaución estas conclusiones, en tanto la muestra utilizada fue reducida. Sólo 20 sujetos respondieron a cada forma de la escala experimental utilizada.

Masters (1974) extiende la investigación realizada por Komorita y Graham (1965). Utiliza para ello dos escalas experimentales diferentes, compuestas ambas por ítems homogéneos[8] y 6 condiciones de opciones de respuesta cada una (2, 3, 4, 5, 6 y 7 alternativas). Para uno de los instrumentos, la fiabilidad de la escala aumentó a medida que se incrementó el número de alternativas de res-

[8] De acuerdo a lo señalado por Komorita y Graham (1965), los ítems serán definidos como homogéneos o heterogéneos en función de sus cargas factoriales. Si estos cargan únicamente sobre un factor, estos autores los definen como homogéneos. Por el contrario, si los ítems que componen la escala cargan sobre dos o más factores, son calificados por estos autores como heterogéneos (Komorita y Graham, 1965).

puesta. Por el contrario, con el otro instrumento no se observaron diferencias significativas en fiabilidad, entre las distintas categorías de respuesta. Esto último permite apoyar solo parcialmente lo sugerido por Komorita y Graham (1965). En su conjunto, estos resultados sugieren que la fiabilidad se ve afectada por el número de alternativas de respuesta en condiciones en las que se observa poca variabilidad en las puntuaciones totales del test, siendo mayor a medida que aumenta el número de categorías de respuesta. "En situaciones donde las opiniones no están demasiado divididas sobre el contenido evaluado, la utilización de un número pequeño de categorías puede repercutir en poca discriminación entre quienes responden, poca variabilidad en la puntuación total y, consecuentemente, en una baja fiabilidad. Cuando se utilizan escalas multicategóricas, la distribución de la puntuación total se dispersa y la fiabilidad incrementa […] el presente estudio ha demostrado que las escalas multicategóricas pueden aumentar la discriminación de los ítems en situaciones con poca varianza" (Masters, 1974, p. 53).

El primer trabajo intensivo con datos simulados es realizado por **Lissitz y Green** (1975). Estos investigadores realizan una simulación del tipo Monte Carlo, en la que analizan la relación entre fiabilidad y el número de opciones de respuesta para una escala tipo Likert. Simulan un test de 10 ítems, aplicado a 50 sujetos. Transforman los resultados en escalas discretas de 2, 3, 5, 7, 9 y 14 puntos (opciones de respuesta). Repiten el proceso con tres niveles diferentes de covariación entre los ítems: .20, .50 y .80., para cada nivel. Tras analizar los resultados, concluyen que la fiabilidad aumenta monotónicamente hasta 5 opciones de respuesta, punto a partir del cual las ganancias en fiabilidad dejan de ser sustantivas. Utilizando una metodología similar, **Jenkins y Taber** (1977) confirman los resultados reportados por Lissitz y Green (1975). Jenkins y Taber (1977) sugieren utilizar entre 5 y 7 opciones de respuesta, pues señalan que so-

bre este número de alternativas de respuesta no se observan ganancias sustanciales en fiabilidad.

Velicer y Stevensons (1978), por su parte, analizan los efectos del número de opciones de respuesta sobre la estructura factorial del *Eysenck Personality Inventory* (EPI), a través de una Análisis de Componentes Principales (ACP). Llevan a cabo un análisis de componentes principales, en el que comparan la estructura factorial que resulta tras utilizar dos y siete opciones de respuesta. En el primer caso (2 opciones de respuesta) se identifican claramente los rasgos de Neuroticismo y Extroversión sugeridos por el EPI (Eysenck y Eysenck, 1969), los cuales dan cuenta de un 18% del total de la varianza. Curiosamente, cuando se utilizan siete opciones de respuesta, la estructura factorial del EPI cambia radicalmente. Aparecen seis componentes, capaces de explicar hasta un 46% de la varianza total, de los cuales un 26% corresponde a los dos primeros, identificados por los autores como *ansiedad general* y *extroversión social*. Los componentes restantes son etiquetados como *compulsividad*, *control de impulsos*, *preocupación por la salud* y *preocupación por cuestiones de relación con terceros*. En su conjunto, estos resultados sugieren que, al menos en lo que al EPI se refiere, el número de opciones de respuesta tiene una clara influencia sobre la validez del instrumento de medición. En tal sentido, los autores sugieren utilizar siete opciones de repuesta por sobre dos, en el entendido que a mayor número de alternativas se observa una clara ganancia respecto del porcentaje de varianza total explicado por el modelo.

Contrariamente a lo señalado por Velicer y Stevensons (1978), **Olsson** (1979) propone que el número de opciones de respuesta no incide sobre la validez del instrumento de medida. Aunque se interesó principalmente en el estudio de los factores que afectan el ajuste de modelos factoriales, las conclusiones a las que llega Olsson (1979) guardan estrecha relación con el estudio del efecto

del número de opciones de respuesta sobre las propiedades psicométricas de los instrumentos para la evaluación de la personalidad. Así, realiza una simulación tipo Monte Carlo en la que se examina el efecto de categorizar variables, con una distribución normal, sobre el ajuste del modelo y la estimación de parámetros en un Análisis Factorial. Aplica como método de extracción de factores *Máxima Verosimilitud*. Genera muestras con distintos tipos de sesgo, al tiempo que varía el número de opciones de repuesta en 2, 3, 4, 5, 7 y 9 alternativas. El principal interés de Olsson (1979) era analizar el grado en que la categorización de variables continuas induce una decisión errónea, en cuanto a extraer más factores de los necesarios. Sugiere que este problema no se observa en situaciones en la que las variables presentan un sesgo de grado similar y en la misma dirección. Por el contrario, cuando las variables están sesgadas en direcciones opuestas, se necesitan utilizar más factores que en el caso anterior para conseguir ajustar el modelo. A pesar de lo señalado, concluye diciendo que el número de opciones de respuesta no incide de forma significativa sobre el ajuste del modelo.

Comrey y Montag (1982) analizan empírica la relación entre el número de opciones de respuesta y la validez factorial. Estudian el efecto del número de opciones de respuesta sobre la estructura factorial del *Comrey Personality Scales* (CPS), por medio de un Análisis de Componentes Principales (ACP). Este cuestionario evalúa siete dimensiones de personalidad. Se compone de 40 ítems, repartidos homogéneamente entre las distintas escalas (5 ítems por escala). Aplican el CPS a muestras independientes, utilizando primero el formato de respuesta tipo Likert original, de 7 alternativas (n= 185 hombres), y luego uno con 2 categorías de respuesta (n= 159 hombres). En su conjunto, los resultados de este estudio indican que las cargas factoriales para las distintas escalas del CPS se muestran, en general, sistemáticamente superiores para la forma con 7 opciones de respuesta, en comparación con la forma de 2 alternativas. Así, los autores

sugieren utilizar este formato de respuesta, antes que el de 2 categorías, a la hora de construir cuestionarios de personalidad.

Aiken (1983) parece confirmar lo señalado años antes por Peabody (1962). Utiliza una escala de 10 ítems para la evaluación del profesorado, la cual varía entre 2 y 7 alternativas de respuesta. Esta es aplicada a un total de 624 estudiantes universitarios que cursan su primer año de carrera. Cada condición experimental es aplicada a sujetos distintos, lo cual resulta en 6 muestras independientes, con cerca de 100 participantes cada una. Los resultados reportados por Aiken (1983) indican que las medias de cada escala crecen linealmente a medida que aumenta el número de opciones de respuesta. Algo similar ocurre con la varianza de cada escala, aunque su crecimiento es curvilíneo. A pesar de ello, la consistencia interna (coeficiente alfa) de la escala se mantiene relativamente constante para las distintas condiciones experimentales evaluadas. Esto lleva a que Aiken (1983) sugiera que el número de opciones de respuesta es independiente de la fiabilidad, por lo que basta con utilizar 2 opciones de respuesta. Asimismo, indica que esta opción es válida, sobre todo, en casos en los que las puntuaciones de la escala muestran un sesgo elevado (apuntamiento y curtosis).

Cicchetti, Showalter y Tyrer (1985) llevan a cabo una simulación Monte Carlo con objeto de analizar el efecto del número de opciones de respuesta (2 a 10 alternativas, más una escala continua de 0 a 100 puntos) sobre la fiabilidad entre jueces. En su conjunto, los resultados reportados por éstos autores sugieren que la fiabilidad aumenta a medida que crece el número de opciones de respuesta. Así, aconsejan utilizar sobre 7 opciones de respuesta. Surgieren que "…las diferencias para la fiabilidad de la escala entre 7, 8, 9 ó 10 categorías ordinales, por un lado, y una escala continua de 100 puntos, por otro lado, es trivial…7 categorías de respuesta aparece como la última funcionalmente inter-

cambiable con una de hasta 100 categorías ordenadas" (Cicchetti *y col*., 1985, p. 35).

McCallum, Keith y Wiebe (1988) analizan el efecto de utilizar 2 ó 6 alternativas de respuesta sobre las propiedades psicométricas del *Multidimensional Health Locus of Control Scales* (MHLCS). Este instrumento consta de 18 ítems, distribuidos de manera homogénea en tres subescalas: *Internal Health Locus of Control* (IHLC), *Chance Health Locus of Control* (CHLC) y *Powerful Others Health Locus of Control* (PHLC). Los autores sugieren que las distintas escalas del MHLCS presentan propiedades psicométricas equivalentes al utilizar 2 ó 6 opciones de respuesta, argumentando a favor de la primera opción (2 alternativas). Aunque nos abstenemos de opinar sobre las propiedades psicométricas del MHLCS reportadas por McCallum *y col.* (1988), una lectura crítica de los resaltados ofrecidos por los autores permite apuntar conclusiones algo diferentes. En primer lugar, McCallum *y col.* (1988) realizan un Análisis de Componentes Principales (ACP) del MHLCS, para las condiciones de 2 y 6 opciones de respuesta. Para ambos formatos de respuesta, los resultados sugieren que los ítems de la escala PHCL cargan sistemáticamente en un solo factor, mientras que para las subescalas IHLC y CHLC, estos cargan en más de un factor. En tal sentido, la escala PHCL puede ser considerada homogénea, en el sentido que todos sus ítems cargan en un solo factor (Komorita y Graham, 1965). Por el contrario, las escalas IHLC y CHLC pueden ser definidas como heterogéneas, en el sentido que sus ítems parecen estar midiendo más de un factor (Komorita y Graham, 1965). Respecto de la consistencia interna, la forma con 6 alternativas de respuesta arroja los siguiente coeficientes alfa: IHLC= .59, CHLC= .67 y PHCL= .76. La forma con 2 opciones muestra los siguientes índices de consistencia interna: IHLC= .42, CHLC= .54 y PHCL= .76. En otras palabras, la forma con 6 opciones de respuesta muestra una fiabilidad sistemáticamente más elevada que la escala con 2 opciones de respuesta, salvo para la escala PHCL, lo cual parece contra-

decir las conclusiones reportadas por McCallum, Keith y Wiebe (1988). En todo caso, estos resultados brindan apoyo a lo sugerido por Komorita y Graham (1965), respecto de que el número de opciones de respuesta parece incidir principalmente sobre escalas con ítems heterogéneos y poco sobre escalas con ítems homogéneos. Así, se observa que la consistencia interna de PHCL no varía en función del número de categorías de respuesta, en tanto que para IHLC y CHLC el valor de alfa crece a medida que aumenta el número de opciones de respuesta.

Oaster (1989) revisa el efecto del número de opciones de respuesta y uso de etiquetas sobre la fiabilidad de cuatro cuestionarios de actitudes diferentes (Oaster, 1984a, 1984b, 1985). En uno de los estudios, presenta a su vez los resultados observados tras realizar un test-retest con formas paralelas de un mismo cuestionario, con un intervalo de 2 a 7 días entre la primera y segunda aplicación. En general, Oaster (1989) analiza el impacto del uso de 3, 5, 7 y 9 alternativas de respuesta sobre la fiabilidad. Las alternativas de respuesta se presentan a los sujetos etiquetadas verbalmente en sus extremos y punto medio. En dos estudios se incluye además una forma de 5 opciones de respuesta en la que se etiqueta cada una de las alternativas de respuesta. En otro estudio se utilizan solo opciones de respuesta con ítems pares y sin respuesta media. En su conjunto, los resultados sugieren sistemáticamente que la consistencia interna de los instrumentos analizados aumenta progresivamente de 3 a 7 opciones de respuesta y comienza a decrecer a partir de 9 alternativas. Evaluada de forma *sui generis*, se observa el mismo patrón respecto a la estabilidad temporal. Nuevamente, esta crece desde 3 a 7 opciones de respuesta y disminuye con 9 categorías.

Sancerini, Meliá y González-Romá (1990) estudian el efecto del formato de respuesta sobre la fiabilidad y validez de criterio del *Cuestionario de Conflicto de Rol* (CCR: Peiró, Meliá, Zacarés y Gonzáles-Romá, 1987). Los participantes en

el estudio fueron 42 trabajadores, con una media de edad de 31 años y diferentes niveles educacionales (42.8%= titulados medios o superiores, 26,2%= bachiller, 19%= formación profesional y 11,9%= estudios primarios). El CCR se compone de 17 ítems, los cuales fueron contestados por los participantes utilizando una escala de 2 ó 4 opciones de respuesta. Con objeto de evaluar su validez de criterio, el CCR fue incluido dentro de una batería de tests con 7 cuestionarios restantes. No se indica el tipo de etiquetas de respuesta utilizado, salvo para el formato dicotómico que fue *verdadero/falso*. Los resultados muestran una consistencia interna (coeficiente alfa: α) similar para ambos formatos de respuesta. En cuanto a la validez, los coeficientes de correlación entre el CCR y las variables criterio se observan en general mayores para el formato de 4 opciones de respuesta. Los autores concluyen señalando que "en su conjunto, los datos sugieren que, al menos por lo que se refiere a este contexto de medición de actitudes y experiencias, las escalas con ítems dicotómicos pueden operar satisfactoriamente con los criterios. Si esto es así, razones de tipo práctico contribuyen a valorar las escalas dicotómicas en el contexto del trabajo profesional" (Sancerni *y col*., 1990a, p. 174).

Por su parte, **Chang** (1994) analiza el efecto del número de opciones de respuesta sobre la fiabilidad y validez de la escala *Quantitative Attitudes Questionnaire* (QAQ), desde una perspectiva Multirasgo-Multimétodo (MRMM) y de Análisis Factorial Confirmatorio (AFC). Trabaja con una muestra de 165 estudiantes de postgrado de la *University of Central Florida*, a los que aplica un cuestionario de 9 ítems (QAQ). Los sujetos responden al QAQ utilizando una escala tipo Likert de 4 ó 6 categorías de respuesta, etiquetadas verbalmente, en dos ocasiones diferentes. El lapso entre aplicaciones es de una semana. Los resultados no permiten decantarse por el uso de uno u otro formato de respuesta. A pesar de lo cual, estos parecen sugerir que si se controla el *método de evaluación*, la escala de 4 opciones de respuesta muestra una estabilidad temporal y consis-

tencia interna mayor que la escala de 6 alternativas. En cuanto a la validez de criterio[9], los resultados parecen indicar que el formato de respuesta no tiene efectos sobre las mismas.

Bandalos y Enders (1996) realizan un estudio de simulación (método Monte Carlo) en el que analizan los efectos del número de opciones de respuesta sobre la fiabilidad del instrumento de medida. Así, simulan un instrumento de 10 ítems, con 100 observaciones. Generan cinco tipos de distribuciones: una Normal (apuntamiento= 0, curtosis= 0), dos distribuciones sesgadas (apuntamiento= 1.75, curtosis= 3.75 y apuntamiento= 2.25, curtosis= 7.0), una platicúrtica (apuntamiento= .25, curtosis= -1.0) y otra leptocúrtica, con apuntalamiento 0 y curtosis= 3.0. Asimismo, analizan las cinco distribuciones bajo tres condiciones de correlación inter-ítem: .25, .50 y .75. Finalmente, las puntuaciones continuas de los ítems se transforman en categoría discretas de 3, 5, 7, 9 y 11 alternativas de respuesta. Se analiza los efectos del diseño previo sobre la fiabilidad de la escala, mediante el coeficiente α. Los resultados de esta investigación sugieren que α tiende a incrementar a medida que la distribución de las escalas se asemeja. A su vez, la fiabilidad se incrementa a medida que aumenta el grado de correlación entre las variables, a pesar de lo cual ésta tiende a nivelarse a partir de 5 ó 7 alternativas de respuesta.

Por su parte, **Ferrando** (2000) propone un Modelo de Ecuaciones Estructurales (SEM) para evaluar la equivalencia de distintos formatos de respuesta, en pruebas de personalidad: (a) binario, (b) tipo Likert de 5, con opciones de respuesta y (c) un continuo de 112 puntos. En lo que a nosotros toca, los resultados ofrecidos por Ferrando (2000) parecen indicar que el ajuste del modelo

[9] De acuerdo con los últimos estándares de la American Psychological Association, la validez de criterio se denomina actualmente *evidencia basada en la relación con otras variables* (American Psychological Association y col., 1999).

factorial, evaluado por medio del *Comparative Fit Index* (CFI), va ganando consistencia a medida que aumenta el numero de opciones de respuesta. Estos resultados son consistentes con un estudio previo, llevado a cabo por el mismo autor (Ferrando, 1995).

Preston y Colman (2000) estudian el efecto del número de opciones de respuesta sobre la fiabilidad, la validez, el poder discriminativo y el tipo formato de respuesta preferido por los evaluados. Participan un total de 149 sujetos, en su mayoría estudiantes universitarios de pregrado de la *University of Leicester*, los cuales son consultados acerca de su parecer sobre los servicios básicos prestados por tiendas o restaurantes a los que han asistido en las últimas semanas. Se crea un cuestionario *ad hoc* con 11 conjuntos de cuestiones, con 5 preguntas cada uno. A cada participante se le presentan las mismas preguntas, variado aleatóriamente las alternativas de respuesta entre 2 y 11 opciones. Cada escala de respuesta muestra solo las etiquetas de los extremos (*muy pobre* y *muy bueno*) y números distribuidos de forma equidistante entre las mismas. Se incluyen además escalas de respuesta de 101 puntos, en las que los sujetos deben marcar sobre un continuo, que va desde 0 (*muy pobre*) hasta 100 (*muy bueno*), su valoración acerca de: a) la calidad global de los servicios y b) su parecer acerca de los distintos formatos de respuesta presentados (2 a 11 opciones de respuesta), respecto de la *facilidad* de uso, *rapidez* para responder y *grado en que permite expresar adecuadamente sus sentimientos*. Se aplica un retest con un lapso entre 1 y 3 semanas entre la primera y segunda aplicación. Los resultados sugieren que los índices de estabilidad temporal más elevados se consiguen con escalas de 7, 8 ,9 ó 10 opciones de respuesta, siendo significativamente más bajos para 2, 3 y 4 alternativas de respuesta. Respecto de la consistencia interna, no se observan diferencias significativas entre los distintos formatos de respuesta, aunque ésta parece incrementar a partir de 7 alternativas de respuesta. La validez de criterio parece ser mayor a partir de 5 opciones de respuesta, aunque no se observan di-

ferencias estadísticamente significativas entre las distintas condiciones presentadas. Finalmente, los sujetos indican que las escalas con 5, 7 ó 10 categorías de respuesta son *fáciles de utilizar*, aquellas con 2, 3 ó 4 son calificadas como *más rápidas* para contestar, pero poco adecuadas para *expresar adecuadamente lo que uno siente*. Sobre esto último, los participantes prefieren las escalas con 10, 11 ó 101 opciones de respuesta. En su conjunto, las escalas con 2, 3 ó 4 alternativas de respuesta aparecen como menos preferidas y las escalas con 10, 9 y 7 categorías, respectivamente, como las más preferidas.

García-Cueto, Muñiz y Lozano (2002) llevan a cabo una simulación (método Monte Carlo) para evaluar el efecto del número de alternativas de respuesta, el tamaño muestral y la correlación entre las variables, sobre la fiabilidad (coeficiente alfa: α), la varianza explicada por el primer factor (Análisis de Componentes Principales) y el ajuste al modelo unifactorial (validez basada en la estructura interna), evaluado mediante el estadístico chi-cuadrado (χ^2). Generan 6400 muestras con distribución normal (0, 1), variando la correlación de 30 variables (0.8 y 0.4) y el tamaño muestral (50, 100, 2000 y 500 casos). Convierten las respuestas en formatos tipo Likert con opciones de respuesta que varían entre 2 y 9 alternativas. Los resultados sugieren que a medida que aumenta el número de alternativas, aumenta también la consistencia interna (coeficiente α), la correlación entre los ítems, el porcentaje de varianza explicado por el primer factor y los valores propios, así como el ajuste χ^2 del modelo unifactorial evaluado. Por otro lado, a medida que disminuye el número de casos, la estructura factorial analizada mediante χ^2 va perdiendo consistencia. En definitiva, sugieren que las propiedades psicométricas de una escala se ven afectadas tanto por el número de alternativas de respuesta, el tamaño muestral, como por la correlación entre los ítems. Sugieren utilizar muestras con un tamaño superior a 100 casos y entre 4 y 9 alternativas de respuesta, siendo 9 opciones el número preferido y 2 el menos recomendado. Un análisis crítico de los resultados reportados

por García-Cueto, Muñiz y Lozano (2002) parecen indicar conclusiones algo diferentes. En primer lugar, para escalas que presentan ítems con una correlación intra-escala elevada (r = .8), el número de opciones de respuesta y el tamaño muestral utilizado parecen no incidir sustantivamente sobre la fiabilidad. Sin importar el tamaño muestral utilizado (50, 100, 200 y 500), se observa un aumento progresivo, aunque no sustantivo, para los coeficientes α que va desde .98 (2 opciones de respuesta) hasta .99 (9 alternativas). Para el caso de la escala con una correlación intra-escala de .4 entre sus ítems, se observa un panorama algo diferente. Nuevamente, el tamaño muestral tiene una mínima incidencia sobre la fiabilidad, no así el número de opciones de respuesta. Los coeficientes α varían progresivamente desde .91 (2 alternativas de respuesta) a .95 (9 opciones). Estos resultados parecen ir en la línea de lo propuesto por Komorita y Graham (1965) y, según se analiza luego, Weng (2004). De acuerdo con estos autores, el número de opciones de respuesta tiene una mayor incidencia sobre escalas con ítems heterogéneos que sobre escalas con ítems homogéneos (Komorita y Graham, 1965; Weng, 2004)[10]. En segundo lugar, la varianza explicada por el primer factor disminuye sistemáticamente a medida que aumenta el tamaño muestral. Si se controla la varianza en función del grado de correlación entre los ítems (r= .8 y r= .4), se observa una disminución general poco sustantiva, salvo para el caso de la escala con ítems con una correlación intra-escala baja (r= .4) con 2 y 3 alternativas de respuesta, donde ésta es más acusada (de 30 a 28 % y de 33 a 27 %, respectivamente, con el porcentaje máximo para n= 100 y el mínimo para n= 500). Cuando se controla el porcentaje de varianza explicada por el primer factor por el número de opciones de respuesta se observa un patrón similar al antes expuesto: por un lado, a medida que aumenta *n*, disminuye el

[10] Tanto Komorita y Graham (1965) como Weng (2004) definen el grado de homogeneidad de los ítems de una escala a partir del tamaño de las cargas factoriales, sobre un factor único. Si los distintos ítems de una escala cargan sobre solo un factor, esta se define como homogénea. Por el contrario, si los ítems de una escala cargan sobre dos o más factores, esta se considera heterogénea (Komorita y Graham, 1965; Weng, 2004).

porcentaje de varianza explicado por el primer factor (aunque no sustantivamente) y, por otro, a medida que aumenta el número de opciones de respuesta, aumenta también el porcentaje de varianza explicado, con un rango que fluctúa entre 46 a 58% (respectivamente, 2 y 9 alternativas de respuesta con *n*= 500). Las fluctuaciones más acusadas se observan para 3 y 4 alternativas de respuesta, en función del tamaño muestral. A pesar de lo señalado, Maydeu-Olivares y Coffman (2005) indican que lo anterior puede deberse a las propiedades de alfa y no al efecto del número de opciones de respuesta propiamente tal. En su conjunto, estos resultados vuelven a dar soporte a las conclusiones de Komorita y Graham (1965) y Weng (2004). Finalmente, es importante recordar que el análisis precedente está hecho sobre datos simulados, por lo que deben ser tomados con precaución. Estos no toman en cuenta el proceso psicológico subyacente por medio del cual los sujetos responden a un cuestionario, por lo que los resultados ofrecidos por García-Cueto, Muñiz y Lozano (2002) deben ser contrastados, necesariamente, con estudios empíricos.

Weng (2004) analiza el impacto de las categorías de respuesta y del tipo de etiquetas utilizado sobre la consistencia interna (coeficiente alfa) y la estabilidad temporal (fiabilidad test-retest) del "*Teacher Attitude Test*" (TeAT). Utiliza dos subescalas del TeAT: *Preocupación por los Otros* (CO) y *Determinación* (DE). La escala CO se compone de 12 ítems heterogéneos[11] que evalúan tanto la *sensibilidad hacia las emociones de los otros* como *la capacidad para compartir las experiencias personales con terceros*. Por su parte, la escala DE contiene 13 ítems, los cuales evalúan de manera homogénea[12] la *determinación para salir adelante en tiempos de frustración y dificultades*. Las cargas factoriales para la escala CO tienen un rango entre 0.43 y 0.74 (M= 0.58, D.S.= 0.09), mientras que para la escala DE entre 0.56 y 0.81 (M= 0.67, D.S.= 0.08). De acuerdo a Weng (2004), el hecho de que DE

[11] Ver nota a pie de página número 10.
[12] Ver nota a pie de página número 10.

muestre cargas factoriales superiores a CO es indicador de que la primera escala resulta más homogénea que CO. Así, Weng (2004) construye 12 condiciones experimentales, compuestas por una combinación de las escalas CO y DE según se indica: a) 6 cuestionarios con cada una de las etiquetas de respuesta claramente identificadas (ALL) y categorías de respuesta entre 3 y 8 alternativas, y b) 6 cuestionarios en los que se indican solo las etiquetas extremas del continuo (END), con categorías de respuesta que varían entre 4 y 9 alternativas. Participan un total de 1247 estudiantes (459 hombre y 788 mujeres) recién ingresados en la *National Taiwan University*. Cada condición experimental es asignada a un grupo diferente, con un rango entre 80 y 121 participantes. Se realiza un retest con un lapso aproximado de 4 semanas, con un rango entre 29 y 43 días de diferencia entre la primera y la segunda aplicación del cuestionario. Los resultados parecen indicar que para la escala CO, el número de opciones de respuesta tiene efectos tanto sobre la consistencia interna (coeficiente α) como sobre la estabilidad temporal (correlación test-retest). Con 7 u 8 alternativas de respuesta, la fiabilidad test-retest para la forma ALL fue mayor que para la forma END. Para el resto de las condiciones, no se observa influencia por parte del tipo de etiquetas utilizado. Por el contrario, para la escala DE no se observan diferencias significativas entre los distintos formatos de respuesta, salvo para la condición de 8 alternativas de respuesta, donde la forma ALL muestra una estabilidad temporal mayor que la forma END. Vistos en su conjunto, estos resultados parecen corroborar las conclusiones de Komorita y Graham (1965) y de García-Cueto, Muñiz y Lozano (2002), respecto de que las escalas heterogéneas se ven más afectadas por el número de opciones de respuesta que las escalas homogéneas. Asimismo, también dan soporte a lo señalado por Masters (1974), quien indica que la fiabilidad de los cuestionarios con mayor variación individual se ve menos afectado por el número de opciones de respuesta utilizado. En este estudio, la escala DE muestra sistemáticamente *desviaciones estándar* más elevadas que la escala CO. Por otro lado, el tipo de etiquetas utilizado parece no afectar la con-

sistencia interna (coeficiente α) de las escalas, tanto para CO como para DE, lo cual es congruente con lo señalado por Churchill y Peter (1984). En todo caso, los resultados parecen indicar que cuando se evalúa la fiabilidad test-retest, identificar claramente cada opción de respuesta con su respectiva etiqueta (forma ALL) aporta mayor estabilidad a las respuestas de los sujetos que el uso etiquetas extremas (forma END). Finalmente, Weng (2004) aconseja utilizar escalas con un mínimo de 5 alternativas de respuesta, ya que a medida que disminuye la dispersión de la escala y aumenta el número de opciones de respuesta, mejoran los índices de fiabilidad. Lo anterior es congruente con lo indicado por Preston y Colman (2000), al tiempo que contradice parcialmente lo sugerido por Lissitz y Green (1975) o Jenkins y Taber (1977). Según estos últimos, no se produce una ganancia sustantiva en fiabilidad sobre 5 alternativas de respuesta. Por el contrario, los resultados reportados por Weng (2004) parecen indicar, al menos para escalas con ítems homogéneos, que la fiabilidad de la escala crece sustantivamente a partir de 5 opciones de respuesta. De todos modos, no se dice nada respecto de cual debe ser el límite superior óptimo de alternativas de respuesta.

Muñiz, García-Cueto y Lozano (2005) estudian el impacto del número de opciones de respuesta sobre las propiedades psicométricas del Eysenck Personality Questionnaire (EPQ), en su versión española (Eysenck y Eysenck, 1978). Específicamente, analizan el efecto de variar el número de categorías de respuesta, entre 2 y 9 opciones, sobre la variabilidad (desviación estándar), fiabilidad (coeficiente alfa) y estructura factorial (Análisis de Componentes Principales) de las tres escalas que componen el EPQ: Neuroticismo (N, 25 ítems), Extroversión (E, 20 ítems) y Psicoticismo (P, 24 ítems). Cada formato de respuesta es aplicado a muestras independientes (Muestra total: 1149 participantes, de los cuales un 571 son mujeres y 578 son hombres). Los resultados reportados indican que la *desviación estándar* de las tres escalas evaluadas (P, E y N) varía en

función del número de opciones de respuesta, incrementando a medida aumenta el número de categorías de respuesta, siendo máximo para 9 (P= 16.7, E= 21.4 y N= 36.8) alternativas y mínimo para 2 opciones de respuesta (P= 5.0, E= 5.3 y N= 7.6). Sobre la fiabilidad, evaluada por medio del *coeficiente alfa* (α), se observa un patrón similar al descrito para el comportamiento de la desviación estándar de las escalas E y N. Así, para las escalas E y N alfa crece en magnitud a medida que aumenta el número de opciones de respuesta, siendo nuevamente máximo para 9 alternativas de respuesta (E= .86 y N= .92) y mínimo para 2 opciones de respuesta (E= .78 y N= .71). A pesar de ello, para la escala P se observa un patrón diferente: alfa crece sistemáticamente a medida que el número de opciones de respuesta varía de 3 hasta 7 alternativas de respuesta (de .64 hasta .83), siendo máximo para éste último formato. En todo caso, para los casos con 2 (α= .77), 8 (α= .72) y 9 (α= .72) alternativas de respuesta se observa un alfa similar y elevado, aunque inferior al obtenido con 7 alternativas de respuesta (α= .83). En lo que toca a la *estructura factorial* (Análisis de Componentes Principales) del EPQ, el porcentaje de varianza explicado por el primer componente de las escala E y N tiende a ser mayor a medida que aumenta el número de opciones de respuesta, siendo máximo para 9 alternativas de respuesta (E= 33% y N= 37%) y mínimo para 2 categorías de respuesta (E= 30% y N= 27%). De todos modos, el incremento de porcentaje de varianza explicada por el primer factor observado, para las escalas E y N, no parece sustantivo. Contrariamente, la escala P presenta un patrón completamente opuesto. Así, el porcentaje de varianza explicado por el primer factor es máximo para 2 opciones de respuesta (35%) y mínimo para 9 alternativas de respuesta (15%). Por último, se analiza el porcentaje de ítems correctamente clasificados, de acuerdo al modelo PEN propuesto por Eysenck (Eysenck y Eysenck, 1978), siendo 5 y 7 alternativas de respuesta los formatos que obtienen un mayor porcentaje de aciertos (entre el 92% y el 100%). En síntesis, los resultados reportados por Muñiz, García-Cueto y Lo-

zano (2005) indican que el número de opciones de respuesta afecta las propiedades psicométricas del EPQ.

En su conjunto, los trabajos revisados hasta este momento han sido realizados desde la óptica de la Teoría Clásica de los Tests (TCT) y/o Análisis Factorial de los Ítems (AFI). Tal y como se analiza a continuación, los trabajos realizados desde la óptica de la Teoría de Respuesta a los Ítems (TRI) son escasos, al menos en lo que se refiere al tema tratado en éste trabajo de Tesis de Doctorado. La primera referencia la encontramos en **Hernández, Muñiz y García-Cueto** (2000), quienes evalúan el modelo de *respuesta graduada* (Muraki, 1990), en función del número de opciones de respuesta. Trabajan con datos simulados y empíricos. Sobre los primeros, se generó una escala de 14 ítems. La simulación de los mismos, sobre la base de un tamaño muestral n=1000, consistió en manipular tanto la dimensionalidad de la escala, como el número de categorías de respuesta de la misma. Así, se generan dos escalas: (a) unidimensional y (b) bidimensional, ambas compuesta por el número de ítems antes indicado. En cuanto a las categorías de respuesta, estas se variaron entre 2 y 9 alternativas. Respecto del estudio empírico (n= 1000), este se basó en el análisis de una escala unidimensional de 14 ítems, diseñada para la evaluación del profesorado. Los sujetos debieron responder a la misma por medio de una escala tipo Likert compuesta por 5 categorías de respuesta. Durante el proceso de análisis de datos, el formato de respuesta de esta última escala fue reagrupado sucesivamente, reduciéndolo de 5 a 4, 3 y 2 alternativas de respuesta. En general, los resultados de ambos estudios (simulación y empírico) parecen sugerir lo siguiente. Al utilizar datos *unidimensionales*, los mejores ajustes se obtienen usando entre cuatro y seis opciones de respuesta, siendo 5 el número de categorías de respuesta sobre el cual se consiguen mejores ajustes del modelo. Por el contrario, con datos *bidimensionales*, los mejores ajustes se consiguen a partir de 6 categorías de respuesta.

Finalmente, **García-Cueto, Muñiz y Lozano** (2003) realizan un trabajo de simulación para estudiar el efecto del número de opciones de respuesta desde la perspectiva de la TRI (*modelo de respuesta graduada de Samejima*). Crean 700 muestras, en las que se simulan la respuesta a 30 ítems, con una intercorrelación de .60. Cada muestra se compone de 500 casos, con escalas de respuesta que varían desde 3 a 9 alternativas. Los resultados reportados sugieren que el parámetro *a* (índice de discriminación) no se ve afectado por el número de opciones utilizado. Por el contrario, a medida que aumenta el número de opciones, incrementa también el grado de *información* aportado por la escala. Sobre este último punto, se observa una paradoja. Aunque la cantidad de información ofrecida por el test aumenta progresivamente desde 4 a 9 opciones de respuesta, curiosamente ésta es máxima para la condición con 3 alternativas. Es más, la cantidad de información ofrecida por esta última condición (3 categorías de respuesta) es incluso superior a la cantidad de información ofrecida por la condición con 9 categorías de respuesta.

Dada la escasez de investigaciones realizadas desde la perspectiva de la TRI, queda claro que es necesario llenar este vacío. Esperamos que el presente trabajo de Tesis de Doctorado contribuya también en este sentido. Por otro lado, la tabla 4 resume las distintas investigaciones revisadas en este apartado, respecto del *estado de la cuestión* de los estudios realizados para examinar el efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios diseñados para evaluar las actitudes y la personalidad.

Tabla 4: Resumen sobre el estado de la cuestión acerca de la evaluación del efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios diseñados para la evaluación de las actitudes y la personalidad.

Autor/es	Tipo de estudio	¿Se observan efectos del número de opciones de respuesta sobre…?		Modelo o Técnica psicométrica utilizada
		Fiabilidad	Validez	
Symonds (1924)	T	SI	--	TCT
Pemberton (1933)	T	SI	--	TCT
Champney y Marshall (1939)	T	SI	--	TCT
Peabody (1962)	E	NO	--	TCT
Komorita y Graham (1965)	E	SI/NO	--	TCT y AFI
Matell y Jacoby (1971)	E	NO	NO	TCT y AFI
Masters (1974)	E	SI/NO	--	TCT
Lissitz y Green (1975)	S	SI	--	TCT
Jenkins y Taber (1977)	S	SI	--	TCT
Velicer y Stevenson (1978)	E	--	SI	ACP
Olsson (1979)	S	--	NO	AFI
Comrey y Montag (1982)	E	--	SI	ACP
Aiken (1983)	E	NO	--	TCT
Cicchetti, Showalter y Tyrer (1985)	S	SI	--	TCT
McCollum, Keith y Wiebe (1988)	E	NO	NO	TCT y ACP
Oaster (1989)	E	SI	--	TCT
Sancerini, Meliá y Gonzalez-Roma (1990)	E	NO	SI	TCT y AFI
Chang (1994)	E	SI	NO	TCT y AFI
Bandalos y Enders (1996)	S	SI	--	TCT
Ferrando (2000)	S	--	SI	AFI (SEM)
Hernández, Muñiz y García-Cueto (2000)	S	--	SI	TRI

(continua en la página siguiente)

(Continuación tabla 4)

Preston y Colman (2000)	E	SI	NO	TCT y AFI
García-Cueto, Muñiz y Lozano (2002)	S	SI	SI	TCT, ACP y AFI
García-Cueto, Muñiz y Lozano (2003)	S	SI/NO	--	TRI
Weng (2004)	E	SI/NO	--	TCT
Muñiz, García-Cueto y Lozano (2005)	E	SI	SI	TCT y ACP

Nota: *Tipo de estudio* = E: empírico, S: simulación y T: teórico; *Modelo o técnica psicométrica utilizada* = AFI: Análisis Factorial de los Ítems, SEM: Modelos de Ecuaciones Estructurales, TCT: Teoría Clásica de los Tests y TRI: Teoría de Respuesta a los Ítems, ACP: Análisis de Componentes Principales.

Sección II: Investigación

Dado a conocer el Marco Teórico (Sección I) que sirve de base para el presente trabajo de Tesis de Doctorado, la **Sección II: Investigación** presenta los estudios empíricos realizados. Se estructura en dos capítulos, los cuales se corresponden con los números cuatro y cinco. Así, el ***capítulo 4*** trata del *método*. En este capítulo se especifican el tipo de *participantes* (apartado 4.1.), los *instrumentos* (apartado 4.2.), los *procedimientos* a realizar (apartado 4.3.) y los *análisis estadísticos* propuestos (apartado 4.4). Por su parte, el ***capítulo 5*** da a conocer los *resultados* obtenidos. En primer lugar, se presentan los *análisis preliminares* efectuados (apartado 5.1). El objetivo de éstos últimos consiste en comprobar si tiene sentido a no llevar a cabo el análisis de datos posterior, según se justifica en el apartado 4.3. Comprobada su viabilidad, los apartados siguientes se centran en la descripción del efecto del número de opciones de respuesta sobre la *fiabilidad* (apartado 5.2) y la *validez* (apartado 5.3) de las escalas *Orientación Negativa hacia los Problemas* (NPO) y *Satisfacción con la Vida* (SWLS). Tanto la fiabilidad como la validez de las escalas NPO y SWLS son analizadas desde la óptica de los modelos de Teoría Clásica de los Test (TCT), Análisis Factorial de los Ítems (AFI) y Teoría de Respuesta a los Ítems (TRI). Específicamente, el estudio de la fiabilidad de los cuestionarios utilizados se estima por medio del coeficiente alfa (modelo de TCT) y el coeficiente omega (modelos de AFI y TRI). Adicionalmente, se analiza la estabilidad temporal de la escala SWLS, bajo los supuestos del modelo TCT. Por su parte, el análisis de la validez de las escalas NPO y SWLS se estima utilizando Modelos de Ecuaciones Estructurales (SEM).

4. *Método*

La presente investigación contempla dos estudios empíricos, identificados con las letras A y B. Según se especifica en el apartado 4.3, el objetivo de realizar dos estudios se fundamenta en la necesidad de poder contar con información que permita contrastar, por un lado, y generalizar, por otro lado, los resultados observados. A continuación se describen ambos estudios.

Estudio A. Durante el año 1992 se realizó un estudio conjunto entre las Universidades de Barcelona e Illinois con objeto de validar las versiones en inglés (Estados Unidos de Norteamérica) y castellano (España) del *Inventario Revisado de Resolución de Problemas Sociales* o *Social Problem Solving Inventory – Revised* (SPSI-R) (D'Zurilla, Nezu y Maydeu-Olivares, 2002; Maydeu-Olivares y D'Zurilla, 1995, 1996; Maydeu-Olivares, Rodríguez-Fornells, Gómez-Benito, J. y D'Zurilla, T. J. 2000). Para tal efecto, se administró una batería de tests compuesta por, entre otros cuestionarios, el *Inventario NEO de los Cinco Factores* (NEO-FFI) y el *Inventario de Afectos Positivos y Afectos Negativos* (PANAS). Estos últimos fueron utilizados para validar el SPSI-R (para más información, consultar D'Zurilla *y col.*, 2002). Complementariamente, con objeto de estudiar las propiedades psicométricas de la escala *Orientación Negativa hacia los Problemas* (NPO) del SPSI-R, durante dicho estudio ésta se aplicó utilizando tres formatos de respuesta diferentes: dos, tres y cinco opciones de respuesta. El análisis de la influencia de tales formatos de respuesta sobre las propiedades psicométricas de la escala NPO no se ha llevado a cabo hasta este momento. Por tal motivo, se utilizan dichos datos como base para realizar el presente *Estudio A*, donde la escala NPO ha sido designada como escala experimental.

Estudio B. Por su parte, los datos para realizar el *Estudio B* fueron recogidos durante el año 2004, entre alumnos de primer ciclo de las carreras de Psicología y Pedagogía, de la Universitat de Barcelona. Los objetivos principales de este segundo estudio son dos: (a) obtener información adicional que permita contrastar las conclusiones sugeridas por el Estudio A y (b) generalizar las mismas más allá del ámbito restringido de este trabajo de Tesis de Doctorado. En consideración al segundo de los objetivos planteados (generalización de las conclusiones), se decide utilizar una escala experimental diferente a la aplicada en el Estudio A. De esta forma, el Estudio B utiliza como escala experimental el cuestionario *Satisfacción con la Vida* (SWLS: Diener, Emmons, Larsen y Griffin, 1985), el cual es aplicado variando su formato de respuesta en dos, tres y cinco opciones. Éste se administra dentro de una batería de tests compuesta por los cuestionarios NEO-FFI, PANAS y SPSI-R (sin la escala NPO).

La elección de las escalas NPO y SWLS como variables experimentales se fundamenta básicamente en cuatro razones: (a) ambas escalas son breves (10 y 5 ítems, respectivamente), (b) ambas escalas son ampliamente conocidas y utilizadas, (c) en tanto constructo psicológico, ambas escalas comparten muchas características (D'Zurilla y Chang, 1995; D'Zurilla y Maydeu-Olivares, 1995; D'Zurilla y Nezu, 1999; D'Zurilla *y col.*, 2002; Diener *y col.*, 1985; Nezu y D'Zurilla, 1989; Pavot y Diener, 1993) y (d), de acuerdo a lo señalado por sus respectivos autores, ambas escalas son unidimensionales y cuentan con adecuadas propiedades psicométricas (para la escala *NPO*, ver p. ej. D'Zurilla *y col.*, 2002; Maydeu-Olivares y D'Zurilla, 1995, 1996; Maydeu-Olivares *y col.*, 2000. Para la escala *SWLS* consultar, entre otros, Arrinddell, Meeuwessen y Huyse, 1991; Diener *y col.*, 1985; Neto, 1993; Pavot y Diener, 1993; Pavot, Diener, Colvin y Sandvik, 1991; Shevlin, Brunsden y Miles, 1998).

Por su parte, se utilizan los cuestionarios NEO-FFI, PANAS y SPSI-R (sin la escala NPO), tanto en el Estudio A como en el Estudio B, con el objeto de contar con variables criterio similares al momento de recoger *evidencias basadas en el grado de relación con otras variables* (American Psychological Association, American Educational Research Association y National Council on Measurement in Education, 1999), dentro del proceso de análisis de la validez de ambas escalas experimentales.

Hechas las aclaraciones previas, a continuación se describen los *participantes* (apartado 4.1), *instrumentos* utilizados (apartado 4.2), *procedimientos* a realizar (apartado 4.3) y *análisis estadísticos* propuestos (apartado 4.4) para los Estudios A y B.

4.1. Participantes

Participaron en este estudio un total 1172 sujetos, todos ellos estudiantes de primer ciclo de las Facultades de Psicología y Pedagogía de la Universitat de Barcelona. La participación de los mismos fue de carácter voluntario.

Los 1172 participantes corresponden a la suma de dos muestras, identificadas como Estudio A y Estudio B.

Estudio A. La muestra del Estudio A consiste en 746 sujetos y está conformada por alumnos de primer ciclo de la carrera de Psicología. La edad de los mismos fluctúa entre los 17 y 53 años, con una media de edad de 20.42 y una desviación estándar (D.S.) de 4.14. En cuanto a la distribución por sexos, un 84.7% fueron mujeres.

Estudio B. Por su parte, la muestra para el Estudio B esta formada por 426 sujetos, alumnos de primer ciclo de las Facultades de Psicología y Pedagogía. La edad de los mismos fluctúa entre 18 y 57 años, con una media de edad de 21.33 años y una D.S. de 4.21. Un 84% de la muestra corresponde a sujetos de sexo femenino.

4.2. Instrumentos

Los cuestionarios utilizados para llevar a cabo la presente investigación fueron los siguientes: (a) el *Inventario NEO de los Cinco Factores* (***NEO-FFI***, Costa y McCrae, 1992, 1999), (b) el *Inventario de Afectos Positivos y Afectos Negativos* (***PANAS***, Watson, Clark y Tellegen, 1988), (c) el *Inventario Revisado de Resolución de Problemas Sociales*, en sus formas Larga y Corta (***SPSI-R***, D'Zurilla, Nezu y Maydeu-Olivares, 2002) y (d) la *Escala de Satisfacción con la Vida* (***SWLS***, Diener, Emmons y Griffin, 1985).

***Inventario NEO de los Cinco Factores* (NEO-FFI).** El *Inventario NEO de los Cinco Factores (*NEO-Five Factor Inventory, en su versión original) es una versión reducida del *Inventario de Personalidad NEO Revisado* (NEO-PI-R: Revised NEO Personality Inventory, Costa y McCrae, 1992, 1999). Basado en el Modelo de los Cinco Grandes o *Five Factor Model* (John, 1990; Widiger y Trull, 1997), el NEO-FFI ha sido diseñado para evaluar cinco dimensiones de personalidad: *Neuroticismo* (N), *Extroversión* (E), *Apertura a la experiencia* (O), *Amabilidad* (A) y *Consciente* [13] (C). Cada escala está compuesta por 12 ítems. Para responder al cuestionario, los sujetos disponen de una escala tipo Likert de cinco opciones de respuesta, por medio de la cual manifiestan su grado de acuerdo o desacuerdo con las distintas preguntas propuestas (0= *Muy en desacuerdo*, 1= *En des-*

[13] En idioma castellano, esta escala recibe distintos nombres: respetuoso, consciente, etc. (Andrés, 1997; Pelechano, 2000; Pervin, 1998).

acuerdo, 2= *Ni de acuerdo ni en desacuerdo*, 3= *De acuerdo*, y 4= *Muy de acuerdo*). Si se desea conocer información adicional sobre las propiedades psicométricas del NEO-FFI, se sugiere consultar Costa y McCrae (1992, 1999).

***Inventario de Afectos Positivos y Afectos Negativos* (PANAS).** El *Inventario de Afectos Positivos y Afectos Negativos* (Positive Affect and Negative Affect Schedule, en su versión original) es un inventario de adjetivos compuesto por dos escalas de 10 ítems cada una, las cuales evalúan dos grandes tipos de estados de ánimo: *Afecto Positivo* (PA) y *Afecto Negativo* (NA). PA refleja el grado en que cada persona siente entusiasmo, energía o alerta (p. ej., energía, concentración, interés, etc.). Por su parte, la escala NA representa una dimensión general ligada a la experimentación subjetiva de sensaciones de angustia o poco placenteras, la cual agrupa estados de ánimo diversos, tales como ira, desprecio, disgusto, etc. (Watson *y col*., 1988). Para completar el cuestionario, los sujetos deben utilizar una escala tipo likert de 5 categorías de respuesta, haciendo referencia al grado en que han experimentado cada una de las emociones planteadas, en un periodo de tiempo determinado (0= *Ligeramente o para nada*, 1= *Un poco*, 2= *Moderadamente*, 3= *Bastante, o* 4= *Mucho*). El PANAS puede ser utilizado para medir los afectos en términos tanto de estado como de rasgo. En nuestro caso, hemos utilizado el formato rasgo, donde los sujetos deben señalar el grado en que *generalmente experimentan las emociones* propuestas. Para más información acerca de la fiabilidad y validez del PANAS, se sugiere consultar Watson *y col*. (1988), Kramp, Maydeu-Olivares, Tous y Gallardo (2004) o Sandin, Chorot, Lostao, Joiner, Santed y Valiente (1999).

***Inventario Revisado de Resolución de Problemas Sociales* (SPSI-R).** El *Inventario Revisado de Resolución de Problemas Sociales* (Social Problem Solving Inventory-Revised, en su versión original) consiste en cinco escalas que evalúan dos dimensiones de *Orientación hacia los Problemas* (Orientación Positiva hacia

los Problemas y Orientación Negativa hacia los Problemas) y tres estilos de *Solución de Problemas* (Estilo Racional de Solución de Problemas, Estilo Impulsivo / Descuidado y Estilo Evitativo). Existen dos formas del SPSI-R, una *larga* (SPSI-RL) y otra *reducida* (SPSI-RS). El SPSI-RL se compone de 52 ítems, distribuidos de la siguiente forma: *Orientación Positiva hacia los Problemas* (PPO = 5 ítems), *Orientación Negativa hacia los Problemas* (NPO = 10 ítems), *Estilo Racional de Solución de Problemas* (RPS = 20 ítems), *Estilo Impulsivo / Descuidado* (ICS = 10 ítems) y *Estilo Evitativo* (AS = 7 ítems). Por su parte, el SPSI-RS es una versión reducida del SPSI-RL. Cada una de sus escalas tiene 5 ítems, siendo 25 el total de ítems para toda la escala. En ambos casos, el cuestionario debe ser contestado con una escala tipo Likert de 5 alternativas de respuesta, por medio de la cual los sujetos indican en qué medida la estrategia de afrontamiento planteada es aplicable o no a su modo característico de enfrentar los problemas de la vida diaria (0= *No se aplica en nada a mi*, 1= *Se aplica levemente a mi*, 2= *Se aplica moderadamente a mi*, 3= *Se aplica mucho a mi* y 4= *Se aplica extremamente a mi*). Si se desea conocer información adicional sobre las propiedades psicométricas del SPSI-R, para las formas americana y española, se sugiere consultar D'Zurilla *y col.* (2002).

***Escala de Satisfacción con la Vida* (SWLS).** La *Escala de Satisfacción con la Vida* (Satisfaction with Life Scale, en su versión original) es una escala compuesta por 12 ítems (Diener *y col.*, 1985), de los cuales cinco se utilizan para evaluar la satisfacción con la vida, mientras que los siete restantes actúan como distractores. En cuanto al constructo *satisfacción por la vida*, éste es definido como la percepción subjetiva sobre los asuntos de la vida diaria, la cual está basada "… en la comparación con los estándares personales de cada individuo, no externamente impuestos" (Diener *y col*., 1985, p. 71). Para la cumplimentación del SWLS, los sujetos utilizan una escala tipo Likert de siete opciones de respuesta, indicando su grado de acuerdo o desacuerdo con las cuestiones propuestas (1= *Muy en desacuerdo*, 2= *En desacuerdo*, 3= *Levemente en desacuerdo*, 4= *Ni de acuerdo*

ni en descuerdo, 5= *Levemente de acuerdo*, 6= *De acuerdo*, y 7= *Fuertemente de acuerdo*). En cuanto a los distractores presentes en el SWLS, se quitaron al azar dos de la escala original, de manera tal de contar con un instrumento compuesto por solo diez ítems. Así, la extensión del SWLS y NPO resulta homogénea. Para más información acerca de la fiabilidad y validez del SWLS, en su forma americana, se sugiere consultar Diener *y col.* (1985). Para la forma española, se sugiere consultar Pívot *y col.* (1991) o Sandin *y col.* (1999).

Hecha la presentación de los instrumentos, cabe señalar que en ambos estudios se utilizaron las adaptaciones al castellano del NEO-FFI y el SPSI-R existentes (para el NEO-FFI consultar Costa y McCrae, 1999; para el SPSI-R ver D'Zurilla *y col.,* 2002). Por su parte, fue necesario adaptar el PANAS (ver anexo 1) y la escala SWLS (ver anexo 2) a dicho idioma. Esto último fue realizado utilizando el *método back-translation*, el cual puede ser definido como un método que permite realizar comparaciones culturales cruzadas de carácter crítico (Berry, 1980). Consta de distintas etapas, las cuales siguen el siguiente orden: (a) un traductor bilingüe traduce el cuestionario en lengua inglesa al castellano (u otro idioma pertinente), (b) otro traductor bilingüe vuelve a traducir la adaptación castellana al Ingles (u otro idioma pertinente) y (c) la prueba original es comparada con el instrumento re-traducido. En esta última etapa se compara cada ítem, buscando eventuales discrepancias entre ambas formas. De hallarse alguna, se continua este proceso hasta conseguir que ambas formas sean semánticamente equivalentes (Brislin, 1980).

Por otro lado, la Tabla 5 muestra las medias $(\overline{X})$ y desviaciones estándar (D.S.) de los cuestionarios NEO-FFI, PANAS, SPSI-R (formas larga y reducida, sin la escala NPO). Para evaluar la *consistencia interna* (fiabilidad) de las distintas pruebas se utilizó el coeficiente *Alfa de Cronbach* (α, Cronbach, 1951). Como muestra la Tabla 5, la consistencia interna de las distintas escalas es en general

buena, aunque para la escala Amabilidad (A) del NEO-FFI ésta puede ser considerada solo aceptable (Estudio A: α= .54; Estudio B: α= .60). En relación a nuestra adaptación al castellano del PANAS, tanto la escala NA (Estudio A: α= .77; Estudio B: α= .85) como la escala PA (Estudio A: α= .74; Estudio B: α= .68) muestran una consistencia interna adecuada. Estos índices de fiabilidad son levemente inferiores a los reportados por Watson *y col*. (1988), para la forma americana. A pesar de ello, son consistentes con las adaptaciones al castellano reportadas por Kramp *y col.* (2004) o Sandin *y col.* (1999).

Tabla 5: Estadísticos descriptivos de las variables criterio, para los Estudios A y B.

Escala	Estudio A			Estudio B		
	Medias	D.S.	α	Medias	D.S.	α
N	24.53	8.23	.86 (.84 ; .88)	23.64	8.37	.85 (.82 ; .88)
E	30.37	6.35	.78 (.75 ; .81)	31.44	6.11	.78 (.73 ; .82)
O	30.58	5.59	.64 (.59 ; .69)	31.32	5.54	.66 (.60 ; .72)
A	30.80	4.74	.54 (.48 ; .60)	32.31	4.83	.60 (.52 ; .67)
C	30.41	6.62	.81 (.78 ; .83)	31.41	6.10	.78 (.74 ; .82)
PA	22.79	5.26	.74 (.70 ; .77)	22.69	4.84	.68 (.61 ; .74)
NA	12.86	6.13	.77 (.74 ; .80)	11.35	6.79	.85 (.82 ; .88)
PPO	12.69 [a]	3.05 [a]	.68 [a] (.63 ; .73)	12.86 [b]	2.67 [b]	.63 [b] (.55 ; .70)
RPS	49.42 [a]	10.16 [a]	.92 [a] (.90 ; .93)	12.77 [b]	2.54 [b]	.67 [b] (.60 ; .74)
ICS	14.29 [a]	5.57 [a]	.84 [a] (.81 ; .86)	6.19 [b]	3.00 [b]	.79 [b] (.74 ; .83)
AS	9.55 [a]	4.81 [a]	.90 [a] (.89 ; .92)	5.74 [b]	2.82 [b]	.81 [b] (.77 ; .85)
	N= 746			N= 426		

Nota. D.S. = Desviación estándar; α = coeficiente alfa, con intervalos de confianza al 99% entre paréntesis; Escalas NEO-FFI: *N* = Neuroticismo, *E* = Extroversión, *O* = Apertura a la experiencia, *A* = Amabilidad, *C* = Consciente; Escalas PANAS: *PA* = Afectos Positivos, *NA* = Afectos Negativos; Escalas SPSI-R [formas Larga (L) y Corta (S)]: *PPO* = Orientación Positiva hacia los Problemas, *RPS* = Estilo Racional de Solución de Problemas, *ICS*= Estilo Impulsivo/Descuidado, *AS*= Estilo Evitativo. N = número de participantes.

[a] SPSI-R, forma Larga.

[b] SPSI-R, forma Corta.

Con objeto de no ser redundantes, las propiedades psicométricas de nuestra adaptación al castellano de la escala SWLS se discuten en el subapartado 5.1.2. En todo caso, se adelanta que éstas se muestran adecuadas, lo cual justifica haber utilizado dicha escala.

4.3. Procedimientos

Cada estudio utilizó una batería de pruebas específica. Para el ***Estudio A***, ésta estuvo compuesta por el NEO-FFI, el SPSI-R (en su versión larga y sin la escala NPO), el PANAS, y la escala NPO. Esta última fue administrada en tres ocasiones distintas, dentro de la batería de pruebas, modificando su formato de respuestas en *dos-*, *tres-* y *cinco-*opciones de respuesta.

La batería de pruebas para el ***Estudio B*** se compuso del NEO-FFI, el SPSI-R (en su versión reducida, sin la escala NPO), el PANAS, y el SWLS. A diferencia de la escala NPO, el SWLS fue administrado en cuatro ocasiones, dentro de la batería de tests. En las primeras tres administraciones, su número de opciones de respuesta varió entre dos, tres y cinco alternativas. La cuarta aplicación fue incorporada para evaluar la estabilidad temporal, intra-sesión, del SWLS. La escala NPO no se utilizó dentro del cuestionario SPSI-R, a fin de de equilibrar las baterías de pruebas de ambas muestras.

Tanto para la escala NPO como para la escala SWLS, las etiquetas utilizadas para evaluar el efecto del número de opciones de respuesta sobre las propiedades psicométricas de ambas escalas fueron los siguientes. Para el formato de ***dos*** opciones de respuesta: 0= SI ó 1= NO. Para el formato de ***tres*** alternativas de respuesta: 0= Falso, 1= A veces verdadero y 2= Verdadero. Finalmen-

te, para el formato de ***cinco*** opciones de respuesta: 0= Muy Falso, 1= Falso, 2= Moderadamente verdadero, 3= Verdadero, 4= Muy verdadero.

Con la intención de asegurar que los participantes no se den cuenta de que están respondiendo el mismo instrumento en más de una ocasión, las pruebas se distribuyeron intercaladamente dentro de cada batería de tests. Para el ***Estudio A,*** el orden de aplicación fue NPO, NEO-FFI, NPO, SPSI-R, NPO y PANAS. Por su parte, el orden de aplicación para el ***Estudio B*** fue SWLS, NEO-FFI, SWLS, SPSI-R, SWLS, PANAS y SWLS, donde la cuarta aplicación del cuestionario SWLS sirve para evaluar la fiabilidad test-retest intra-sesión de dicho instrumento. Para controlar el efecto de los distintos formatos de respuesta, se utilizó un diseño parcialmente contrabalanceado. Así, se construyen tres baterías de pruebas para cada instrumento experimental (NPO y SWLS). La Tabla 6 muestra el orden de aparición de los distintos formatos de respuesta, incluido el retest para SWLS, en cada batería de tests. Finalmente, las baterías de pruebas fueron distribuidas al azar entre los participantes, aunque cuidando equilibrar su número dentro de cada muestra (ver Tabla 6).

Tabla 6: Orden de presentación de los distintos formatos de respuesta para las escalas NPO y SWLS, en función de la batería de tests aplicada.

Batería de tests	Orden de los diferentes formatos de respuesta	Test-retest para SWLS	Estudio A	Estudio B
			N	N
A	5 – 3 – 2	5	249 (33.34%)	140 (32.86%)
B	3 – 2 – 5	3	251 (33.65%)	147 (34.51%)
C	2 – 5 – 3	2	246 (33.01%)	139 (32.63%)
		TOTAL	746 (100%)	426 (100%)

En definitiva, la batería de pruebas utilizada en ambas muestras difiere tan solo en el instrumento experimental utilizado. Mientras que en el *Estudio A*

se aplica la escala **NPO** (ver anexo 3), en el *Estudio B* se utiliza la escala **SWLS** (ver anexo 2). A nuestro juicio, el hecho de utilizar dos instrumentos experimentales diferentes permite examinar si los resultados observados son específicos a un instrumento en particular o no (NPO o SWLS). Asimismo, dentro del Estudio B se analiza la estabilidad temporal, intra-sesión, de la escala SWLS.

Finalmente, los sujetos completaron su respectiva batería de pruebas en una sesión que duró aproximadamente una hora, luego de recibir una explicación general respecto de los fines y contenidos de la evaluación, así como del modo como debían de cumplimentar los cuestionarios. Es importante destacar que se trabajó exclusivamente con sujetos voluntarios.

4.4. Análisis estadísticos

Los análisis estadísticos se realizan en tres etapas, tanto para la escala NPO como para la escala SWLS, cada una de las cuales incluye un análisis de datos desde la perspectiva de tres modelos psicométricos diferentes: *Teoría clásica de los Tests* (TCT), *Análisis Factorial de los Ítems* (AFI) y *Teoría de Respuesta a los Ítems* (TRI). La **primera etapa** o **análisis preliminares** consiste en examinar la correlación que se establece entre los distintos formatos de respuesta (dos, tres y cinco alternativas de respuesta). Así, para el ***modelo TCT*** se analiza el grado de correlación entre las distintas *puntuaciones directas* o *totales* estimadas (T), tras utilizar dos, tres y cinco opciones de respuesta. Para el ***modelo AFI*** se examina el grado de correlación entre los distintos *factores* estimados (η), tras utilizar dos, tres y cinco alternativas de respuesta. Finalmente, Para el ***modelo TRI*** se evalúa el grado de correlación entre los distintos *rasgos latentes* estimados (θ), tras utilizar dos, tres y cinco categorías de respuesta.

La **segunda etapa** o **análisis de la fiabilidad** se centra en evaluar el efecto de variar el número de opciones de respuesta en dos, tres y cinco alternativas, sobre la *consistencia interna* de las escalas NPO y SWLS, bajo la óptica de los tres modelos psicométricos antes mencionados (TCT, AFI y TRI). Adicionalmente y solo desde la perspectiva del modelo de TCT, se examina la *estabilidad temporal* de la escala SWLS.

Finalmente, en la **tercera etapa** o **análisis de la validez** se examina el efecto de variar el número de opciones de respuesta en dos, tres y cinco alternativas, sobre la *evidencia basada en el grado de relación con otras variables*, desde la perspectiva de los modelos de TCT, AFI y TRI. Adicionalmente, bajo los supuestos de los modelos de AFI y TRI se examina la *evidencia basada en la estructura interna*, en función de los distintos formatos de respuesta propuestos.

La mayor parte de los análisis estadísticos se realizan utilizando el programa Mplus v.3.13 (L. K. Muthén y Muthén, 2005). Los *coeficientes alfa* (α) y *omega* (ω), junto a sus respectivos intervalos de confianza, son calculados por medio del programa Mathematica v.5.0.1 (Wolfram Research, 2003), siguiendo a Yuan, Guarnaccia y Hayslip (2003) y Maydeu-Olivares, Coffman y Hartmann (2006). Todos los análisis se realizan sin supuestos de normalidad. Para estimar los modelos propuestos dentro de **TCT** y **AFI**, se utiliza como estimador *máxima verosimilitud, con errores estándar robustos a la no normalidad* y *pruebas de bondad de ajuste corregidas por su media asintótica* (Satorra y Bentler, 1994) [en su conjunto, identificado con las siglas ***MLM*** (L. K. Muthén y Muthén, 2004)]. Por su parte, el modelo propuesto dentro de **TRI** se estima utilizando un enfoque de *Modelos de Ecuaciones Estructurales* (SEM), con un estimador de *mínimos cuadrados con una matriz de pesos diagonal con errores estándar robustos* y la *prueba de bondad de ajuste de Satorra y Bentler* (en su conjunto, identificado con las siglas ***WLSM*** (Muthén, 1993; Muthén *y col.*, 1997; L. K. Muthén y Muthén, 2004)]. Para comparar los

modelos anidados propuestos, tanto dentro del modelo AFI como TRI, se utiliza el la prueba c*hi-cuadrado escalado de Satorra-Bentler*, sin supuestos de normalidad, propuesta por Satorra y Bentler (2001) y descrita por B.O. Muthén y Muthén (2005).

A continuación se ofrece una descripción detallada de los distintos análisis de datos propuestos para cada una de las tres etapas de *análisis estadísticos* antes enunciadas, en función del modelo psicométrico utilizado. Así, el subapartado 4.4.1 trata el análisis de datos desde el modelo de Teoría Clásica de los Tests (TCT), el subapartado 4.4.2 describe el análisis de datos desde la perspectiva del modelo de Análisis Factorial de los Ítems (AFI) y, finalmente, el subapartado 4.4.3 versa sobre el análisis de datos desde la óptica del modelo de Teoría de Respuesta a los Ítems (TRI).

4.4.1 Teoría Clásica de los Tests

En este apartado se describen en detalle los procedimientos de ***análisis estadísticos*** propuestos para analizar el efecto del número de opciones de respuesta sobre las propiedades psicométricas de las escalas NPO y SWLS, bajo los supuestos del modelo de *Teoría Clásica de los Tests* (TCT). Consta de tres subapartados, los cuales describen en detalle las distintas etapas de análisis de datos enunciadas anteriormente. Así, el subapartado 4.4.1.1 o *primera etapa* trata los análisis preliminares. El subapartado 4.4.1.2 o *segunda etapa* aborda el análisis de la fiabilidad y, finalmente, el subapartado 4.4.1.3 o *tercera etapa* indica como se procede con el análisis de la validez de las escalas experimentales utilizadas en este trabajo de Tesis de Doctorado.

4.4.1.1 Primera etapa.

La ***primera etapa*** o ***análisis preliminares*** consiste en examinar la correlación que se establece entre las puntuaciones directas o totales estimadas (*T*), tanto para la escala NPO como para la escala SWLS, tras variar el formato de respuesta en dos, tres y cinco alternativas. Así, para el ***Estudio A***, se calcula *T* para cada una de las versiones experimentales de la escala **NPO** (*dos-*, *tres-* y *cinco*-opciones de respuesta). Luego se correlacionan las distintas *T* estimadas entre sí. Si las correlaciones entre las diferentes versiones experimentales de la escala NPO son significativamente iguales a uno, se asume que el número de opciones de respuesta no tiene efectos sobre las propiedades psicométricas de ésta escala. De observarse tal hecho, no se prosigue con el análisis de datos propuestos para las siguientes etapas. Por el contrario, si las correlaciones son distintas de uno, se asume que pueden existir diferencias sobre la fiabilidad o la validez de la escala NPO, cuando se varía el formato de respuesta en *dos-*, *tres-* y *cinco*-categorías de respuesta. De darse este último caso, se prosigue con el análisis de datos posterior, según se especifica más adelante.

En el ***Estudio B*** se realiza el mismo procedimiento descrito anteriormente, excepto que la escala experimental utilizada es **SWLS**.

4.4.1.2 Segunda etapa.

La ***segunda etapa*** o a***nálisis de la fiabilidad*** consiste en evaluar el efecto de variar el número de opciones de respuesta en dos, tres y cinco alternativas, sobre la *consistencia interna* de las escalas NPO y SWLS. Adicionalmente, se examina la *estabilidad temporal* intra-sesión de la escala SWLS, para las distintas condiciones experimentales (dos, tres y cinco opciones de respuesta). Específicamente, en el ***Estudio A*** se calcula la consistencia interna para cada una de las

condiciones experimentales de la escala **NPO** (*dos-*, *tres-*, y *cinco-* opciones de respuesta). Ésta es estimada utilizando el *coeficiente alfa* (α: Cronbach, 1951). Asimismo se calculan intervalos de confianza (IC) al 99% para cada uno de los coeficientes alfa estimados, siguiendo a Yuan, Guarnaccia y Hayslip (2003) y Maydeu-Olivares, Coffman y Hartmann (2006). Los IC se utilizan para comparar la significación estadística de las diferencias observadas entre los coeficientes alfa estimados para cada una de las condiciones experimentales consideradas. Así, si el valor estimado de α de una condición experimental cae dentro de los límites del intervalo de confianza de otra condición, se concluye que la condición en cuestión no difiere significativamente de la otra. De observarse este hecho, se concluye que no existe efecto del número de opciones de respuesta sobre la consistencia interna de la escala NPO, entre las condiciones experimentales con IC solapados. Por el contrario, si el valor estimado de α para una condición cae fuera del intervalo de confianza de las condiciones experimentales restantes, se concluye que dicha condición es significativamente diferente de las otras. De darse este último caso, se asume que existe efecto del número de opciones de respuesta sobre la consistencia interna de la escala NPO, entre las condiciones experimentales con un α significativamente diferente.

En el ***Estudio B*** se lleva acabo el mismo procedimiento descrito anteriormente, excepto que la escala experimental utilizada es **SWLS**. Adicionalmente, se examina la estabilidad temporal, intra sesión, de las distintas versiones experimentales de la escala SWLS (*dos-*, *tres-*, y *cinco-* opciones de respuesta). En este último caso, se estiman las puntuaciones directas o totales (T) para la primera (test) y segunda (retest) aplicación de las versiones experimentales de la escala SWLS, junto a sus respectivos intervalos de confianza (IC) al 99%. Las puntuaciones T para el test y retest de cada condición experimental son correlacionadas entre sí (entre el tests-retest de la condición con *dos* opciones de respuesta, entre el tests-retest de la condición con *tres* alternativas de respuesta

y entre el tests-retest de la condición con *cinco* categorías de respuesta). Luego se examinan los IC *entre* condiciones experimentales (entre *dos-*, *tres-*, y *cinco-*opciones de respuesta), de manera tal de poder comparar la significación estadística de las diferencias observadas para la estimación test-retest entre las condiciones experimentales consideradas. Así, si el valor estimado de test-retest de una condición experimental cae dentro de los límites del IC de otra condición, se concluye que la condición en cuestión no difiere significativamente de la otra. De observarse este hecho, se concluye que no existe efecto del número de opciones de respuesta sobre la estabilidad temporal de la escala SWLS, entre las condiciones experimentales con IC solapados. Por el contrario, si el valor estimado de test-retest para una condición cae fuera del IC de las condiciones experimentales restantes, se concluye que dicha condición es significativamente diferente de las otras. De darse este último caso, se asume que existe efecto del número de opciones de respuesta sobre la estabilidad temporal de la escala SWLS, entre las condiciones experimentales con un valor test-retest estimado significativamente diferente.

4.4.1.3 Tercera etapa.

La ***tercera etapa*** o ***análisis de la Validez*** se centra en examinar el efecto de variar el número de opciones de respuesta en dos, tres y cinco alternativas, sobre la *evidencia basada en la relación con otras variables* de las escalas NPO y SWLS. En el ***Estudio A***, se evalúa en qué medida la correlación entre las puntuaciones directas o totales estimadas (*T*) de las distintas condiciones experimentales de la escala NPO (*dos-*, *tres-*, y *cinco*-opciones de respuesta) y las variables externas (escalas NEO-FFI, PANAS y SPSI-R, donde ésta última no incluye la escala NPO) se mantiene invariante a medida que incrementa el número de alternativas de respuesta. El objetivo de este análisis consiste en determinar si existe algún tipo de efecto al variar el número de opciones de respuesta sobre la

evidencia basada en el grado de relación con otras variables de la escala NPO. Para ello se propone un Modelo de Ecuaciones Estructurales (SEM), el cual se ilustra en la figura 2.

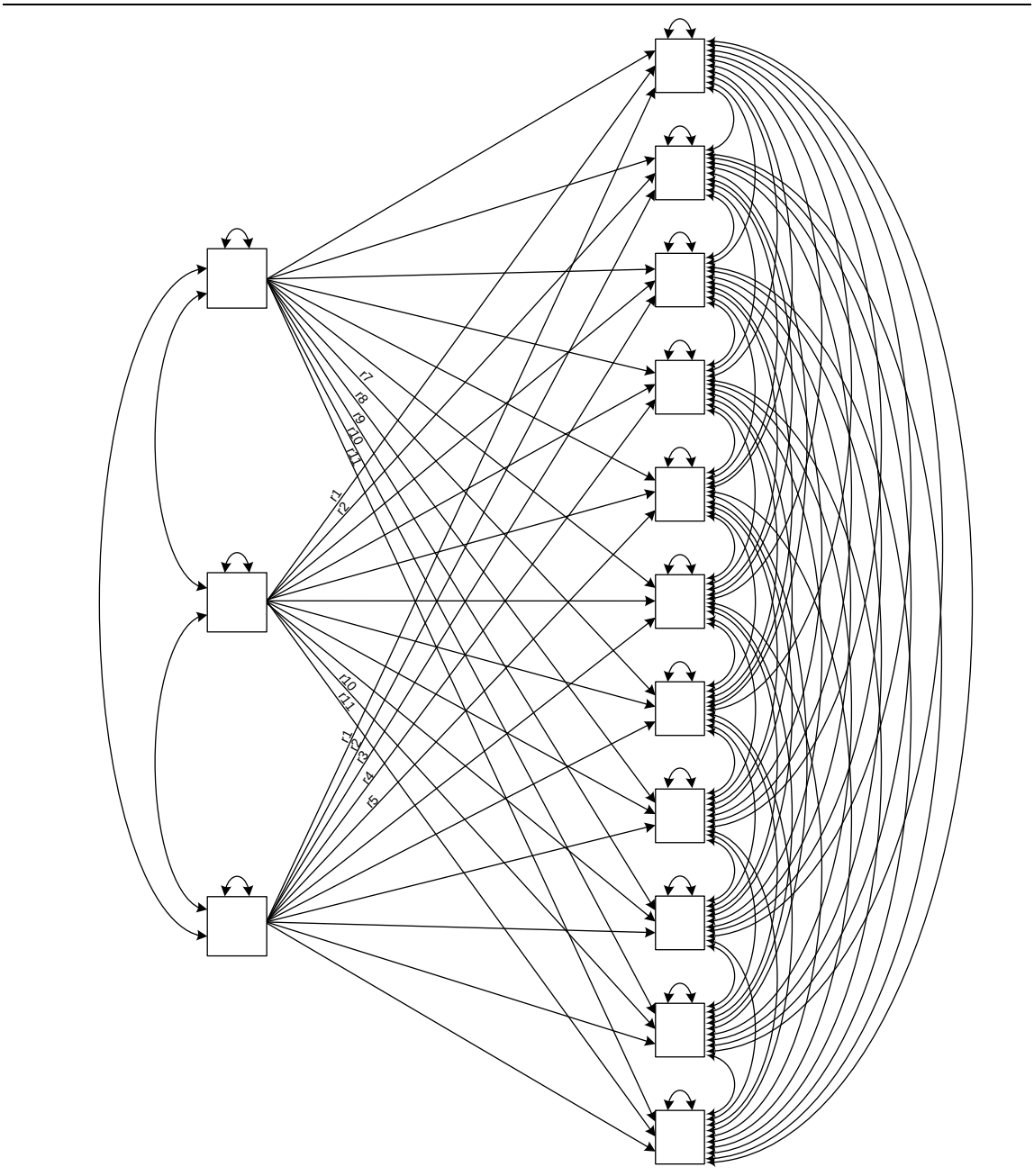


Figura 2: Modelo de Ecuaciones Estructurales (SEM) para evaluar la relación de la escala NPO con variables externas, bajo el modelo de Teoría Clásica de los Tests (TCT).

Nota: Ejemplo de Modelo de Ecuaciones Estructurales (SEM) para evaluar la *validez* (*evidencia basada en el grado de relación con otras variables*) de la escala NPO, para los distintitos formatos de respuesta, bajo el modelo de Teoría Clásica de los Test (TCT). Se utiliza el mismo modelo para evaluar la escala SWLS. Escalas NEO-FFI: *N* = Neuroticismo, *E* = Extroversión, *O* = Apertura a la experiencia, *A* = Amabilidad, *C* = Consciente; Escalas PANAS: *PA* = Afectos Positivos, *NA* = Afectos Negativos; Escalas SPSI-R [formas Larga (L) y Corta (S)]: *PPO* = Orientación Positiva hacia los Problemas, *RPS* = Estilo Racional de Solución de Problemas, *ICS*= Estilo Impulsivo/Descuidado, *AS*= Estilo Evitativo.

Si se analiza la figura 2, se observa que aparecen en su margen izquierdo *tres* variables observadas. Estas variables observadas se corresponden con las puntuaciones directas o totales estimadas (*T*) al utilizar *dos-*, *tres-* y *cinco-* opciones de respuesta para la escala **NPO**. Denominamos ***NPO-2op*** a la puntuación total (*T*) obtenida tras sumar las puntuaciones directas de cada ítem de la escala experimental NPO con *dos* alternativas de respuesta. Se identifica como ***NPO-3op*** a la puntuación total (*T*) obtenida tras sumar las puntuaciones directas de cada ítem de la escala experimental NPO con *tres* categorías de respuesta. Finalmente, denominamos ***NPO-5op*** a la puntuación total (*T*) obtenida tras sumar las puntuaciones directas de cada ítem de la escala experimental NPO con *cinco* alternativas de respuesta.

Por otro lado, en el margen derecho de la figura 2 aparecen las puntuaciones totales (*T*) de las distintas variables observadas externas utilizadas para evaluar la *evidencia basada en el grado de relación con otras variables* de la escala NPO. Se ha denominado *variables criterio* a las variables observadas que aparecen en el margen derecho de la figura 2. En sentido estricto, la relación de las variables criterio entre sí no es de interés dentro del modelo propuesto, salvo para evaluar la *evidencia basada en el grado de relación con otras variables* de la escala NPO. Por tal motivo, no se establecen restricciones sobre la relación entre variables criterio. De esta forma, según se aprecia en la figura 2, todas las variables criterio aparecen correlacionadas entre sí. Por el contrario, las escalas NPO-2op, NPO-3op y NPO-5op aparecen correlacionadas entre sí, ya que las tres miden el mismo constructo.

En definitiva, las únicas restricciones que se incluyen en el modelo de la figura 2 se encuentran en la relación de las variables NPO-2op, NPO-3op y NPO-5op entre sí con las variables criterio. Por ejemplo, nótese que la relación entre NPO-2op, NPO-3op y NPO-5op con la variable criterio Neuroticismo (*N*),

del *Inventario NEO de los Cinco Factores* (NEO-FFI), se asume como constante. Igualmente, la relación entre NPO-2op, NPO-3op y NPO-5op con la variable criterio Estilo Evitativo (*AS*), del *Inventario de Resolución de Problemas Sociales* (SPSI-R), se asume como constante.

Por lo tanto, la estimación de la bondad de ajuste del modelo SEM propuesto en la figura 2 permite evaluar si la relación entre las condiciones experimentales NPO-2op, NPO-3op y NPO-5op con las variables criterio es constante o no. De acuerdo a los criterios para valorar la bondad de ajuste absoluta y aproximada que se describen más adelante, si el modelo presentado en la figura 2 muestra un ajuste adecuado (absoluto y/o aproximado), se concluye que el número de opciones de respuesta no afecta la evidencia basada en el grado de relación con otras variables de la escala NPO. Por el contrario, si el modelo sugerido en la figura 2 no ajusta de forma absoluta y/o aproximada, ello implica que existen diferencias significativas en las correlaciones entre las puntuaciones directas o totales estimadas (*T*) de alguna condición experimental (NPO-2op, NPO-3op y NPO-5op) con alguna/s variable/s criterio (N, E, O, A, C, PA, NA, PPO, RPS, ICS o AS).

De no observarse ajuste (absoluto y/o aproximado) en el modelo propuesto en la figura 2, se estiman adicionalmente intervalos de confianza (IC) al 99% para cada una de las correlaciones, entre las variables observadas y las variables criterio. Los distintos IC estimados son examinados uno a uno con objeto de comparar la significación estadística de cada estimación. Así, si el valor estimado para la correlación entre una condición experimental (NPO-2op, NPO-3op y NPO-5op) y alguna variable criterio (N, E, O, A, C, PA, NA, PPO, RPS, ICS o AS) cae dentro de los límites del intervalo de confianza de otra condición, se concluye que la condición en cuestión no difiere significativamente de la otra. De observarse este hecho, se concluye que no existe efecto del número de op-

ciones de respuesta sobre la evidencia basada en el grado de relación con otras variables de la escala NPO, entre las condiciones experimentales con IC solapados. Por el contrario, si el valor estimado para la correlación entre una condición experimental (NPO-2op, NPO-3op y NPO-5op) y alguna variable criterio (N, E, O, A, C, PA, NA, PPO, RPS, ICS o AS) cae fuera de los límites del intervalo de confianza de otra condición, se concluye que la condición en cuestión si difiere significativamente de la otra. De darse este último caso, se asume que el número de opciones de respuesta afecta la evidencia basada en el grado de relación con otras variables de la escala NPO, entre las condiciones experimentales con un valor de correlación estimado significativamente diferente.

Respecto del ***Estudio B***, se utiliza el mismo procedimiento descrito en los párrafos precedentes, aunque reemplazando la escala NPO por la escala SWLS. Por tanto, en este caso se denomina ***SWLS-2op*** a la puntuación total (*T*) obtenida tras sumar las puntuaciones directas de cada ítem de la escala experimental SWLS con *dos* alternativas de respuesta. Se identifica como ***SWLS-3op*** a la puntuación total (*T*) obtenida tras sumar las puntuaciones directas de cada ítem de la escala experimental SWLS con *tres* categorías de respuesta. Finalmente, se denomina ***SWLS-5op*** a la puntuación total (*T*) obtenida tras sumar las puntuaciones directas de cada ítem de la escala experimental SWLS con *cinco* alternativas de respuesta.

Según se indicó en el capítulo 2, subapartado 2.4.1, para valorar la **bondad de ajuste absoluta** del modelo restringido propuesto en la figura 2, tanto para la escala NPO como para la escala SWLS, se utiliza el índice chi-cuadrado (X^2), sin supuestos de normalidad. Específicamente, se estima el *índice* X^2 *con media corregida de Satorra-Bentler* (Satorra y Bentler, 1994). Adicionalmente, se estiman los índices de **bondad de ajuste aproximada** RMSEA, TLI y CFI, los cuales son valorados en función de los criterios expuestos en el capítulo 2 (subapar-

tado 2.4.1)[14]. Finalmente, se calculan intervalos de confianza al 99% para RMSEA, por medio del programa FITMOD (Browne, 1992; Browne y Cudeck, 1993).

4.4.2 Análisis Factorial de los Ítems

El apartado 4.4.2 describe en detalle los procedimientos de ***análisis estadísticos*** propuestos para analizar el efecto del número de opciones de respuesta sobre las propiedades psicométricas de las escalas NPO y SWLS, bajo los supuestos del modelo de *Análisis Factorial de los Ítems* (AFI). Consta de tres subapartados, los cuales describen en detalle las distintas etapas de análisis de datos enunciadas anteriormente. Así, el subapartado 4.4.2.1 o *primera etapa* trata los análisis preliminares. El subapartado 4.4.2.2 o *segunda etapa* aborda el análisis de la fiabilidad y, finalmente, el subapartado 4.4.2.3 o *tercera etapa* indica como se procede con el análisis de la validez de las escalas experimentales utilizadas en este trabajo de Tesis de Doctorado.

4.4.2.1 Primera etapa.

La ***primera etapa*** o ***análisis preliminares*** consiste en examinar la correlación que se establece entre los factores estimados (η), tanto para la escala NPO como para la escala SWLS, tras variar el formato de respuesta en *dos*, *tres* y *cinco* alternativas, de acuerdo al modelo que se propone en la figura 3. Antes de describir en detalle los elementos que componen esta figura, conviene señalar que

[14] Según se indicó en el capítulo 2 (subapartado 2.4.1), para ***RMSEA*** se considera *adecuado* un valor comprendido entre 0 y 0.05, mientras que *aceptable* un valor que oscile entre 0.05 y 0.08 (Browne y Cudeck, 1993). Por su parte, para los índices ***TLI*** y ***CFI*** se consideran *adecuados* valores por sobre un umbral de 0.9, siendo 1.00 reflejo de un ajuste perfecto (Bentler, 1990; Hair y *col.*, 1999; Tucker y Lewis, 1973).

la misma representa un ejemplo de Análisis Factorial Confirmatorio, aplicable tanto bajo los supuestos del modelo AFI como bajo los supuestos del modelo de Teoría de Respuesta a los Ítems (TRI). Lo que sigue se refiere principalmente al primero de los modelos antes mencionados (AFI).

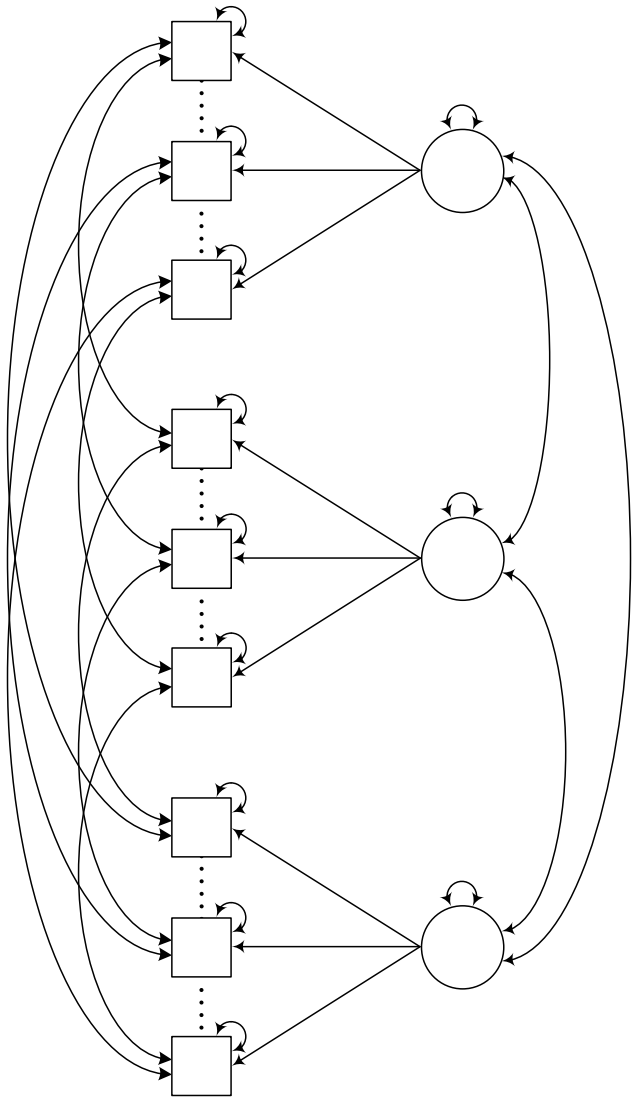


Figura 3: Modelo de Ecuaciones Estructurales (SEM) para la evaluación de la correlación entre factores y validez estructural de la escala NPO, bajo los supuestos de los modelos de Análisis Factorial de los Ítems (AFI) y Teoría de Respuesta a los Ítems (TRI).

Nota. Ejemplo de Modelos de Ecuaciones Estructurales (SEM) para la evaluación de la correlación entre factores y la validez estructural de los distintos formatos de respuesta, bajo los modelos de Análisis Factorial de los Ítems (AFI) y Teoría de Respuesta a los Ítems (TRI), para la escala NPO. Desde el modelo de AFI, los datos son tratados como continuos. Por el contrario, desde el modelo TRI éstos son tratados como discretos. En el entendido que la escala NPO consta de 10 ítems, las líneas punteadas representan una abstracción de los ítems no representados directamente en la figura. Se utiliza el mismo modelo para evaluar la escala SWLS.

De acuerdo con la figura 3, para el ***Estudio A*** se estiman tres factores (NPO-2op, NPO-3op y NPO-5op). Estos son estimados a partir de los ítems que dan cuenta de cada una de las versiones experimentales de la escala NPO, los cuales están representados en el margen izquierdo de la figura 3. Nótese que las líneas punteadas que aparecen en dicha figura sirven de abstracción para representar los diez ítems que componen cada una de las versiones experimentales de la escala NPO. Así, los diez ítems de la escala NPO con 2 opciones de respuesta sirven de indicadores para estimar el factor NPO-2op, los diez ítems de la escala NPO con 3 alternativas de respuesta sirven de indicadores para estimar el factor NPO-3op y los diez ítems de la escala NPO con 5 categorías de respuesta sirven para estimar el factor NPO-5op.

Como se aprecia en la figura 3, los tres factores estimados (NPO-2op, NPO-3op y NPO-5op) aparecen correlacionados entre sí, por cuanto miden el mismo constructo. Asimismo, nótese también que los residuos de los ítems cuyo enunciado es el mismo se hallan correlacionados entre sí (ver margen izquierdo de la figura 3). Así, por ejemplo, el residuo del ítem 1 de la escala NPO cuando se utilizan 2 opciones (NPO-2op) se halla correlacionado con el residuo del ítem 1 cuando se utilizan 3 (NPO-3op) y 5 (NPO-5op) alternativas de respuesta. A su vez, el residuo del ítem 1 cuando se utilizan 3 opciones (NPO-3op) se halla correlacionado con el residuo del ítem 1 cuando se utilizan 5 categorías (NPO-5op) de respuesta. Si no se incluye dentro del modelo propuesto la restricción respecto a que deben correlacionar entre sí los residuos de aquellos ítems con un mismo enunciado, se incurre en un grave error de especificación (*mispecation*). Esto se debe a que la estimación de las correlaciones entre los ítems idénticos, pero con distinto número de opciones de respuesta, son el resultado de dos efectos: (a) el efecto de que todos los ítems evalúan el mismo constructo y (b) el efecto de utilizar el mismo enunciado, en ciertas correlaciones. Gracias a las especificaciones realizadas en el modelo representado en la fi-

gura 3, se separan ambos efectos. Por lo tanto, esto permite evaluar si variar el número de opciones de respuesta, en *dos-*, *tres-* y *cinco-*opciones, afecta o no las propiedades psicométricas de la escala NPO.

En definitiva, el objetivo del análisis de datos para esta etapa consiste en examinar si las correlaciones entre los factores estimados (η) son estadísticamente iguales a uno. Adicionalmente, se calculan intervalos de confianza al 99% para las correlaciones entre los η. Si las correlaciones entre los distintos η son significativamente iguales a uno, se concluye que el número de opciones de respuesta no afecta las propiedades psicométricas de la escala NPO, bajo el modelo de AFI. Si este fuese el caso, no se procede con el análisis de datos posterior. Por el contrario, si las correlaciones entre los η son distintas de uno, ello sugiere que puede existir efecto del número de opciones de respuesta sobre la fiabilidad o validez de la escala NPO, cuando se utiliza un formato de respuesta u otro. De darse este último caso, se prosigue con el análisis de datos sugerido en las siguientes tapas.

4.4.2.2 Segunda etapa.

La ***segunda etapa*** o a***nálisis de la fiabilidad*** consiste en evaluar la *consistencia interna* de las escalas NPO y SWLS, para cada una de las condiciones experimentales propuestas. Específicamente, en el ***Estudio A*** se calcula la consistencia interna para cada una de las condiciones experimentales de la escala **NPO** (*dos-*, *tres-*, y *cinco-* opciones de respuesta). En este caso, ésta es estimada utilizando el *coeficiente omega* (ω: McDonald, 1999). Asimismo, se calculan intervalos de confianza (IC) al 99% para cada uno de los coeficientes ω estimado, como una extensión al procedimiento propuesto por Yuan, Guarnaccia y Hayslip (2003) y Maydeu-Olivares, Coffman y Hartmann (2006). Los IC se utilizan la

comparar la significación estadística de las diferencias observadas entre los coeficientes ω estimados. Así, si el valor estimado del coeficiente ω de una condición experimental cae dentro de los límites del intervalo de confianza de otra condición, se concluye que la condición en cuestión no difiere significativamente de la otra. De observarse este hecho, se concluye que no existe efecto del número de opciones de respuesta sobre la consistencia interna de la escala NPO, entre las condiciones experimentales con IC solapados. Por el contrario, si el valor estimado del coeficiente ω para una condición cae fuera del intervalo de confianza de las condiciones experimentales restantes, se concluye que dicha condición es significativamente diferente de las otras. De darse este último caso, se asume que existe efecto del número de opciones de respuesta sobre la consistencia interna de la escala NPO, entre las condiciones experimentales con un coeficiente ω significativamente diferente.

En el ***Estudio B*** se lleva acabo el mismo procedimiento de análisis de datos descrito anteriormente, excepto que la escala experimental utilizada es **SWLS**.

4.4.2.3 Tercera etapa.

La ***tercera etapa*** o ***análisis de la Validez*** sirve para analizar si existe efecto del número de opciones de respuesta (1) sobre la relación con otras variables (*evidencia basada en el grado de relación con otras variables*) y/o (2) sobre la estructura interna (*evidencia basada en la estructura interna*) de las escalas NPO y SWLS, bajo el modelo de Análisis Factorial de los Ítems (AFI).

Para evaluar si el incremento del número de opciones de respuesta afecta la **evidencia basada en el grado de relación con otras variables**, se procede de

la siguiente manera. Se construyen dos modelos SEM, uno restringido y otro sin restricciones (libre). El *modelo restringido* combina el modelo factorial ilustrado en la figura 3 con el patrón de correlaciones entre las condiciones experimentales y las variables criterio (N, E, O, A, C, PA, NA, PPO, RPS, ICS o AS) que aparece en la figura 2. Este modelo aparece representado gráficamente en la figura 4. Por su parte, el *modelo sin restricciones* (libre) sirve de contraste.

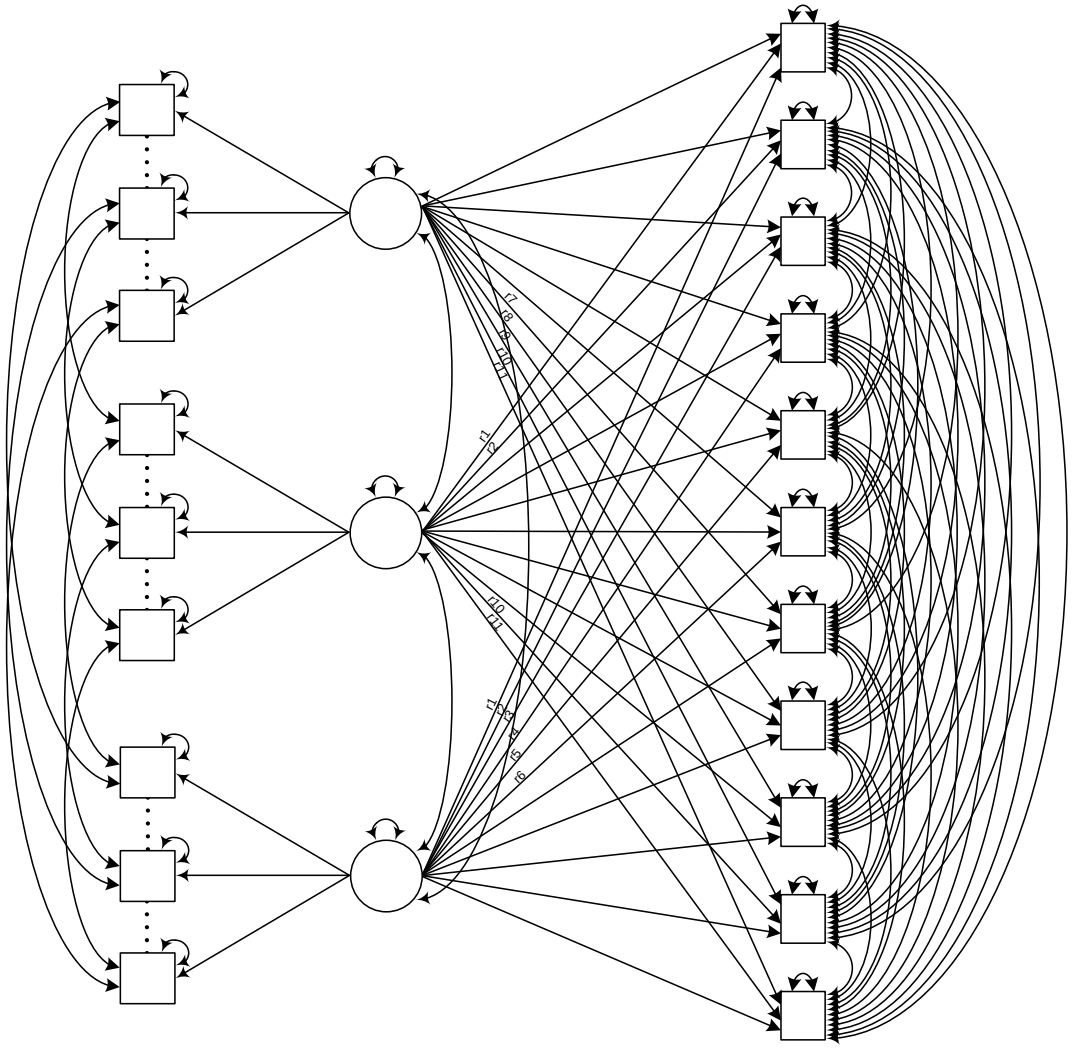


Figura 4: Modelo de Ecuaciones Estructurales (SEM) para evaluar *evidencia basada en el grado de relación con otras variables* de la escala NPO, bajo los supuestos de los modelos de Análisis Factorial de los Ítems (AFI) y Teoría de Respuesta a los Ítems (TRI).

Nota: Ejemplo de Modelo de Ecuaciones Estructurales (SEM) para evaluar la *evidencia basada en el grado de relación con otras variables* de la escala NPO, para los distintitos formatos de respuesta, bajo los modelos de Análisis Factorial de los Ítems (AFI) y Teoría de Respuesta a los Ítems (TRI). En el entendido que la escala NPO consta de 10 ítems, las líneas punteadas representan una abstracción de los ítems no representados directamente en la figura. Se utiliza el mismo modelo para evaluar la escala SWLS. Escalas NEO-FFI: *N* = Neuroticismo, *E* = Extroversión, *O* = Apertura a la experiencia, *A* = Amabilidad, *C* = Consciente; Escalas PANAS: *PA* = Afectos Positivos, *NA* = Afectos Negativos; Escalas SPSI-R [formas Larga (L) y Corta (S)]: *PPO* = Orientación Positiva hacia los Problemas, *RPS* = Estilo Racional de Solución de Problemas, *ICS*= Estilo Impulsivo/Descuidado, *AS*= Estilo Evitativo.

Tomando como referencia la figura 4, la diferencia entre los modelos libre y restringido radica en que para éste último se indica que las correlaciones entre los factores estimados (η), derivados a partir de las distintas condiciones experimentales (*dos*, *tres* y *cinco* opciones de respuesta), y las variables criterio (N, E, O, A, C, PA, NA, PPO, RPS, ICS o AS) se especifican para que sean iguales entre sí. Por ejemplo, de acuerdo a lo observado en la figura 4, la relación entre el factor NPO-2op con AS, la relación entre el factor NPO-3op con AS, y la relación entre el factor NPO-5op con AS es igual al coeficiente r_{11}. Por el contrario, en el modelo libre las relaciones entre NPO-2op con AS, NPO-3op con AS, y NPO-5op con AS son diferentes entre sí, en tanto pueden variar libremente.

En el entendido que el modelo restringido propuesto en la figura 4 está anidado dentro del modelo libre, se puede evaluar la diferencia entre ambos modelos por medio de la prueba *chi-cuadrado escalado de Satorra-Bentler*, sin supuestos de normalidad (*Satorra-Bentler Scaled Chi-Square*: Satorra y Bentler, 2001). En el caso de no observarse una diferencia estadísticamente significativa entre ambos modelos, ello implica la relación entre la escala NPO y las variables criterio no se ve afectada por el número de opciones de respuesta. Por el contrario, de apreciarse una diferencia estadísticamente significativa entre ambos modelos, el modelo libre debe preferirse al modelo con restricciones. A su vez, ello es indicador de que el número de opciones de respuesta podría afectar *evidencia basada en el grado de relación con otras variables* de la escala NPO. De ser este el caso, se estiman adicionalmente intervalos de confianza (IC) al 99% para cada uno de los parámetros r_x que relacionan los factores NPO-2op, NPO-3op y NPO 5 op con las variables criterio. Los distintos IC estimados son examinados uno a uno, con objeto de comparar la significación estadística de cada estimación. Así, si el valor estimado para la correlación entre una condición experimental (NPO-2op, NPO-3op y NPO-5op) y alguna variable criterio (N, E, O, A, C, PA, NA, PPO, RPS, ICS o AS) cae dentro de los límites del intervalo de confianza de otra con-

dición, se concluye que la condición en cuestión no difiere significativamente de la otra. De observarse este hecho, se concluye que no existe efecto del número de opciones de respuesta sobre la evidencia basada en el grado de relación con otras variables de la escala NPO, entre las condiciones experimentales con IC solapados. Por el contrario, si el valor estimado para la correlación entre una condición experimental (NPO-2op, NPO-3op y NPO-5op) y alguna variable criterio (N, E, O, A, C, PA, NA, PPO, RPS, ICS o AS) cae fuera de los límites del intervalo de confianza de otra condición, se concluye que la condición en cuestión difiere significativamente de la otra. De observarse este último caso, se asume que el número de opciones de respuesta si afecta la evidencia basada en el grado de relación con otras variables de la escala NPO, entre las condiciones experimentales con un valor de correlación estimado significativamente diferente.

En cuanto al ***Estudio B***, se realiza el mismo procedimiento de análisis de datos descrito anteriormente, aunque en este caso se utiliza la escala **SWLS** como variable experimental.

Respecto del análisis del efecto del número de opciones de respuesta sobre la **evidencia basada en la estructura interna** de las escalas NPO y SWLS, se construyen una serie de modelos unifactoriales para cada condición experimental (*dos*, *tres* y *cinco* opciones de respuesta) y cuestionario diana (escala NPO o escala SWLS). Así, en el ***Estudio A*** se analizan tres modelos unifactoriales, los cuales se corresponden con las tres condiciones experimentales de la escala NPO (NPO-2op, NPO-3op y NPO-5op). La figura 5 muestra un ejemplo de modelo unifactorial para NPO con 2 opciones de respuesta. De acuerdo con dicha figura, el factor NPO-2op se estima a partir de los ítems que dan cuenta de la versión experimental de la escala NPO con dos opciones de respuesta, los cuales están representados en el margen izquierdo de la figura 5. Así, los diez ítems de la escala NPO con 2 opciones de respuesta sirven de indicadores para esti-

mar el factor NPO-2op. Se aplica el mismo modelo para estimar los factores de las condiciones experimentales de la escala NPO restantes. De esta forma, los diez ítems de la escala NPO con 3 opciones de respuesta sirven de indicadores para estimar el factor NPO-3op, en tanto que los diez ítems de la escala NPO con 5 opciones de respuesta sirven de indicadores para estimar el factor NPO-5op.

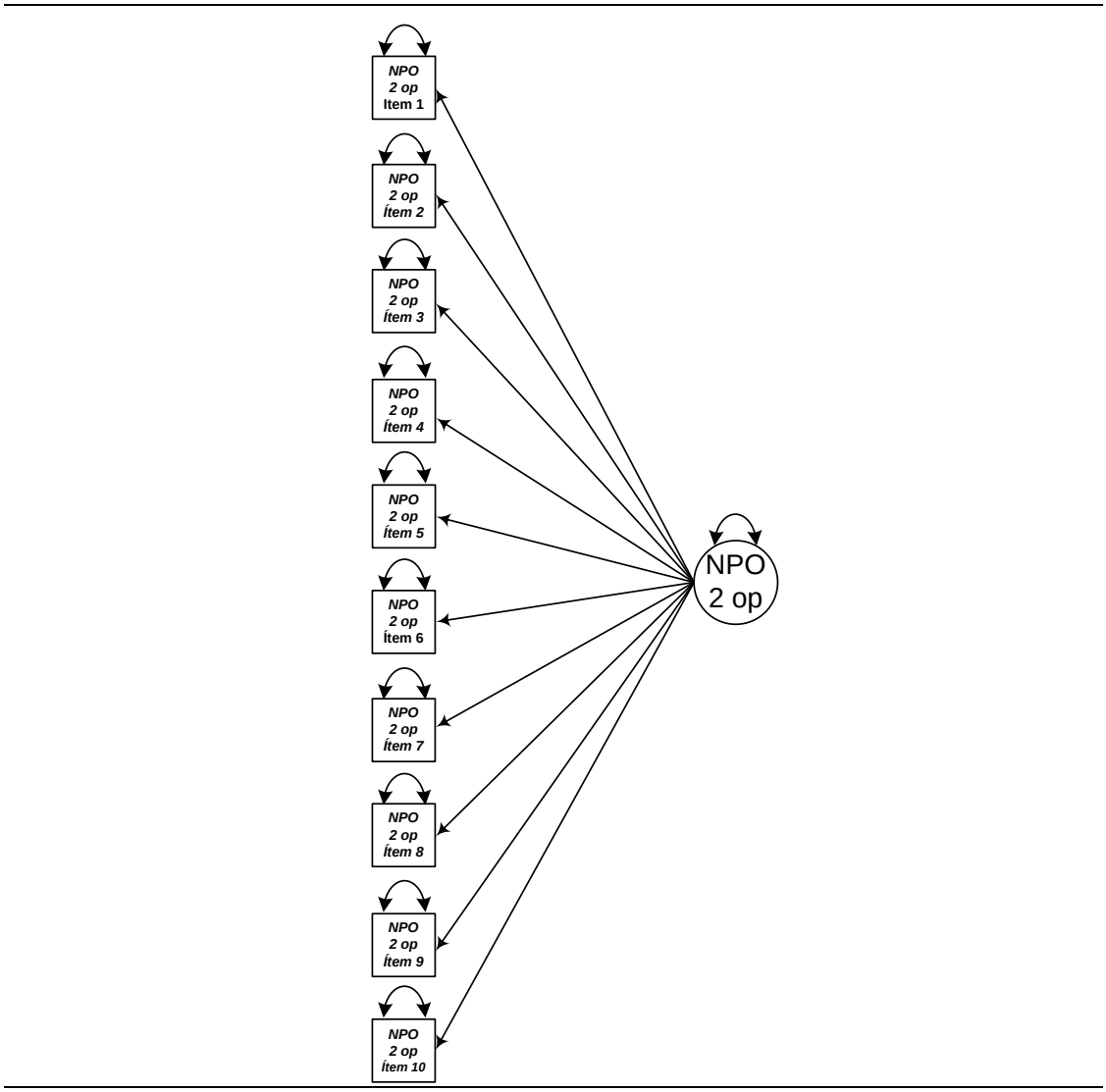


Figura 5: Modelo SEM para evaluar *la evidencia basada en la estructura interna* de la escala NPO, para los distintos formatos de respuesta, bajo los supuestos de los modelos de Análisis Factorial de los Ítems (AFI) y Teoría de Respuesta a los Ítems (TRI).

Nota: ejemplo de modelos de ecuaciones estructurales (SEM) para evaluar la *evidencia basada en la estructura interna* de la escala NPO, bajo los modelos de Análisis Factorial de los Ítems (AFI) y Teoría de Respuesta a los Ítems (TRI). Dentro del modelo AFI los datos son tratados de forma continua, mientras que bajo los supuestos del modelo TRI estos son tratados como discretos. Se utiliza el mismo modelo para analizar la escala SWLS.

Una vez estimados los modelos unifactoriales para cada condición experimental de la escala NPO (NPO-2op, NPO-3op y NPO-5op), se compara la bondad de ajuste absoluta y/o aproximada observada *entre* condiciones experimentales (NPO-2op, NPO-3op y NPO-5op), de acuerdo a los criterios que se exponen más adelante. El objetivo de este análisis consiste en comprobar si la bondad de ajuste absoluta y/o aproximada observada *entre* condiciones experimentales se mantiene constante o varía a medida que cambia el número de opciones de respuesta. En el caso de que la bondad de ajuste absoluta y/o aproximada observada entre condiciones experimentales (NPO-2op, NPO-3op y NPO-5op) se mantenga *constante*, se concluye que el número de opciones de respuesta no afecta la evidencia basada en la estructura interna de la escala NPO. Por el contrario, si la bondad de ajuste absoluta y/o aproximada observada *varía* entre condiciones experimentales (NPO-2op, NPO-3op y NPO-5op), se concluye que el número de opciones de respuesta si afecta la evidencia basada en la estructura interna de la escala NPO.

Por su parte, en cuanto al ***Estudio B*** se aplica el mismo procedimiento de análisis de datos descrito anteriormente, aunque en este caso se utiliza la escala **SWLS** como variable experimental. Así, los diez ítems de la escala SWLS con 2 opciones de respuesta sirven de indicadores para estimar el factor SWLS-2op. Los diez ítems de la escala SWLS con 3 opciones de respuesta sirven de indicadores para estimar el factor SWLS-3op, mientras que los diez ítems de la escala SWLS con 5 opciones de respuesta sirven de indicadores para estimar el factor SWLS-5op.

Finalmente, de acuerdo a los criterios expuestos en el capítulo 2 (subapartado 2.4.1), la **bondad de ajuste absoluta** del modelo SEM unifactorial propuesto en la figura 5, tanto para la escala NPO como para la escala SWLS, se valora a partir del índice chi-cuadrado (X^2), sin supuestos de normalidad. Especí-

ficamente, se estima el *índice X^2 con media corregida de Satorra-Bentler* (Satorra y Bentler, 1994). Adicionalmente, se estiman los índices de **bondad de ajuste aproximada** RMSEA, TLI y CFI, los cuales son valorados en función de los criterios señalados en el capítulo 2 (subapartado 2.4.1)[15]. Finalmente, se calculan intervalos de confianza al 99% para RMSEA, por medio del programa FITMOD (Browne, 1992; Browne y Cudeck, 1993).

4.4.3 Teoría de Respuesta a los Ítems

El apartado 4.4.3 describe los procedimientos de ***análisis estadísticos*** propuestos para analizar el efecto del número de opciones de respuesta sobre las propiedades psicométricas de las escalas NPO y SWLS, bajo los supuestos del modelo de *Teoría de Respuesta a los Ítems* (TRI). Los análisis estadísticos propuestos en esta sección no difieren mayormente de los descritos en el apartado 4.4.2, donde se trató el modelo de Análisis Factorial de los Ítems (AFI). Por tanto, se puede consultar la sección antes indicada para profundizar lo que se expone a continuación.

De acuerdo a los señalado en los capítulos 2.3 y 2.4, los distintos procedimientos de análisis de datos, bajo los supuestos del modelo de TRI, se basan en el *modelo de respuesta graduada de Samejima* (1969). Lo anterior se fundamenta, a su vez, en el procedimiento descrito por Muthén (Muthén, 1984), el cual se encuentra implementado en el programa Mplus (L. K. Muthén y Muthén, 2005).

---

[15] Según se indicó en el capítulo 2 (subapartado 2.4.1), para ***RMSEA*** se considera *adecuado* un valor comprendido entre 0 y 0.05, mientras que *aceptable* un valor que oscile entre 0.05 y 0.08 (Browne y Cudeck, 1993). Por su parte, para los índices ***TLI*** y ***CFI*** se consideran *adecuados* los valores que están por sobre un umbral de 0.9, siendo 1.00 reflejo de un ajuste perfecto (Bentler, 1990; Hair y *col.*, 1999; Tucker y Lewis, 1973).

Desde la perspectiva SEM aquí utilizada, la diferencia principal entre el modelo TRI y el modelo AFI es que en el primer caso los datos son tratados como categóricos, mientras que en el segundo éstos son tratados como continuos. Así, los análisis estadísticos presentados anteriormente para el modelo AFI (apartado 4.4.2) se realizan a partir de la matriz de covarianzas entre los ítems y las variables criterio. Por el contrario, los análisis estadísticos propuestos bajo el modelo TRI (modelo de respuesta graduada de Samejima) se modelizan a partir de una matriz de correlaciones Policóricas, de Pearson y Poliseriales, tal y como se describe a continuación.

De acuerdo a lo señalado por B.O. Muthén (1984), el modelo SEM de TRI aquí utilizado se estima a partir de una matriz de correlaciones que consta de (a) correlaciones *Policóricas* entre los ítems categóricos, (b) correlaciones *de Pearson* entre las variables criterio (pues estas son continuas), y (c) correlaciones *Poliseriales* entre los ítems categóricos y las variables criterio continuas. Una vez estimada esta matriz de correlaciones, el resto de las etapas de análisis estadísticos propuestos en esta sección sigue los mismos criterios expuestos en el apartado 4.4.2, para el modelo de AFI.

De esta forma, en la ***primera etapa*** se realizan los *análisis preliminares*, siguiendo los mismos criterios expuestos en el subapartado 4.4.2.1. La ***segunda etapa*** se centra en el análisis de la *fiabilidad*, donde se evalúa la consistencia interna de las escalas experimentales (NPO y SWLS), en función de los distintos formatos de respuesta (*dos*, *tres* y *cinco* opciones de respuesta), a través del coeficiente omega. Se evalúa el efecto de las distintas condiciones experimentales sobre la fiabilidad de las escala NPO y SWLS según los criterios expuestos en el subapartado 4.4.2.2. Finalmente, la ***tercera etapa*** se centra en el análisis del efecto del número de opciones de respuesta sobre la estructura interna (*evidencia basada en la estructura interna*) y la relación con otras variables (*evidencia basada en el*

grado de relación con otras variables) de las escalas NPO y SWLS, de acuerdo a los criterios expuestos en el subapartado 4.4.2.3.

5. Resultados

El capítulo 5 presenta los resultados observados, de acuerdo a los criterios expuestos en el subapartado 4.4. Así, se exponen primero los resultados desde la óptica de la *Teoría Clásica de los Tests* (TCT), luego desde la perspectiva del *Análisis Factorial de los Ítems* (AFI) y, por último, desde el modelo de *Teoría de Respuesta a los Ítems* (TRI). En cada caso, se procede a estimar cada una de las etapas de análisis estadísticos propuestos en el subapartado 4.4.

5.1 Teoría Clásica de los Tests

A continuación se ofrecen los resultados observados, tras analizar las escalas NPO y SWLS, de acuerdo a los criterios expuestos en el subapartado 4.4.1. Así, en primer lugar se presentan los *análisis preliminares* (subapartado 5.1.1), luego el *análisis de la fiabilidad* (subapartado 5.1.2) y, finalmente, el *análisis de la validez* (subapartado 5.1.3), desde la óptica de la Teoría Clásica de los Tests (TCT).

5.1.1 Primera etapa: Análisis preliminares

La correlación entre las puntuaciones directas (T) de los distintos formatos de respuesta de las escalas NPO y SWLS, junto a sus respectivos intervalos de confianza al 99%, se presentan en la tabla 7. Según se observa en dicha tabla, la correlación entre las distintas condiciones experimentales (*dos*, *tres* y *cinco* opciones de respuesta) de la escala NPO varía entre .84 (NPO-2op) y .87 (NPO-5op), en tanto que para la escala SWLS éstas están comprendidas en un rango que va desde .79 (SWLS-2op) hasta .86 (SWLS-5op). En ninguno caso se observa

que el límite superior de los intervalos de confianza alcance un valor de .90 o mayor. En tal sentido, tanto para la escala NPO como para la escala SWLS, el patrón de correlaciones observado entre las distintas condiciones experimentales parece indicar que el hecho de variar el número de opciones de respuesta puede afectar las propiedades psicométricas de los instrumentos de evaluación en estudio. Por tal motivo, parece oportuno continuar con los análisis estadísticos posteriores. En todo caso, conviene tener presente que, aunque distinto de 1, la elevada magnitud de las correlaciones observadas entre los distintos formatos de respuesta (*T*), tanto para la escala NPO y como para la escala SWLS, permiten inferir que el efecto del número de opciones de respuesta sobre las propiedades psicométricas de los instrumentos utilizados en esta investigación puede ser moderado o bajo.

Tabla 7: Correlación entre los diferentes tipos de formatos de respuesta para las escalas NPO y SWLS, desde el modelo de Teoría Clásica de los Tests (TCT).

Formatos de Respuesta relacionados	NPO		SWLS	
	ρ	*IC*	ρ	*IC*
2 y 3	.85**	(.82 ; .88)	.79**	(.74 ; .84)
2 y 5	.84**	(.82 ; .87)	.81**	(.77 ; .85)
3 y 5	.87**	(85 ; .89)	.86**	(.83 ; .89)

Nota: ρ = Coeficiente de correlación; *IC* = Intervalo de confianza al 99 % para ρ. TCT = Teoría Clásica de los Tests. NPO = Orientación Negativa hacia los Problemas; SWLS = Escala de Satisfacción con la Vida.
** $p < .01$.

Quizás, la falta de correlación perfecta entre las puntuaciones directas o totales (*T*) de las escalas NPO y SWLS, cuando se varía el número de opciones de respuesta, puede deberse a una baja estabilidad temporal. Para analizar esta hipótesis, se evalúa la fiabilidad test-retest intra-sesión de la escala SWLS, en función de las diferentes condiciones experimentales (dos, tres y cinco opciones de respuesta). Los resultados estimados para la estabilidad temporal intra-

sesión de la escala SWLS aparecen en la tabla 8. Como se observa en dicha tabla, la correlación test-retest, aunque elevada, es distinta de uno. Para la condición experimental de 2 opciones de respuesta (SWLS-2op), ésta es de .88; para la condición de 3 alternativas (SWLS-3op) es de .89 y para la condición de 5 opciones de respuesta (SWLS-5op) es de .91. Por su parte, el examen de los intervalos de confianza (IC) estimados para las distintas las condiciones experimentales revela que estos se solapan entre sí en todos los casos. Así, la comparación de la estimación puntual para la fiabilidad test-retest entre condiciones experimentales de la escala SWLS no revela diferencias estadísticamente significativas. Por otro lado, las correlaciones test-retest reportadas en la tabla 8 son superiores a las observadas en la Tabla 7, para la relación entre los distintos formatos de respuesta de la escala SWLS.

Tabla 8: Estabilidad temporal para la escala SWLS.

Formato de respuesta	N	r test-retest
2	139	.88** (.82 - .94)
3	146	.89** (.85 - .94)
5	140	.91** (.88 - .95)

Nota: entre paréntesis aparecen los intervalos de confianza al 99% para los coeficientes de correlación test-retest (r test-retest).
** $p < .01$.

Así, de acuerdo a lo señalado, es posible inferir que la falta de correlación perfecta entre las distintas condiciones experimentales de la escala SWLS, aunque elevada, se debe no solo a la influencia de variar el número de opciones de respuesta, sino también al hecho de que los sujetos parecen mostrar cierto grado de inconsistencia al responder dicha escala en dos ocasiones diferentes.

5.1.2 Segunda etapa: Análisis de la fiabilidad

La tabla 9 muestra los valores estimados para la *consistencia interna* de las escalas NPO y SWLS, bajo los supuestos del modelo de Teoría Clásica de los Test (TCT). Como se observa en dicha tabla, tanto para la escala NPO como para la escala SWLS, la magnitud *de la estimación puntual* del coeficiente alfa (α) se incrementa a medida que aumenta el número de opciones de respuesta. Asimismo, el examen de los intervalos de confianza (IC) para las distintas condiciones experimentales indica que estos se encuentran solapados en la mayor parte de los casos observados. A pesar de ello, en el entendido que la estimación puntual de muchos de los casos examinados no se encuentra dentro del IC de otra condición experimental, se puede concluir que existen diferencias estadísticamente significativas para la mayor parte de los casos analizados.

Tabla 9: Consistencia Interna para las escalas NPO y SWLS, bajo los supuestos del modelo de Teoría Clásica de los Test (TCT).

Número de opciones de respuesta	α	
	NPO	SWLS
2	.78 (.75 ; .81)	.62 (.56 ; .69)
3	.84 (.81 ; .86)	.71 (.65 ; .76)
5	.88 (.86 ; .90)	.76 (.71 ; .81)

Nota: α = Coeficiente Alfa de Cronbach. Entre paréntesis se incluyen los intervalos de confianza al 99% para α.

Respecto de la escala **NPO**, el valor de la estimación puntual del coeficiente α incrementa desde .78 (NPO-2op) hasta .88 (NPO-5op). Por otro lado, los IC estimados se solapan en todos los casos analizados. De todos modos, la estimación puntual de cada condición experimental no se halla dentro de los límites del IC de las condiciones restantes. Así, el incremento de α, a medida que

aumenta el número de opciones de respuesta, puede ser considerado estadísticamente significativo para todos los casos analizados.

Por su parte, para la escala **SWLS** se observan resultados similares a los antes expuestos. En tal sentido, la tabla 9 muestra que el valor *de la estimación puntual* de α vuelve a incrementar a medida que aumenta el número de opciones de respuesta. En este caso, desde .62 (SWLS-2op) hasta .76 (SWLS-5op). Asimismo, los IC estimados se solapan nuevamente en todos los casos analizados. De todos modos, a diferencia de la escala NPO, la estimación puntual de la condición experimental SWLS-5op se halla dentro de los límites del IC de la condición experimental SWLS-3op. Así, el incremento de α, a medida que aumenta el número de opciones de respuesta, puede ser considerado estadísticamente significativo solo para la relación entre las condiciones experimentales SWLS-2op y SWLS-3op o para la relación entre SWLS-2op y SWLS 5 op.

En su conjunto, los resultados observados parecen indicar que el número de opciones de respuesta afecta de forma leve o moderada la expresión de la consistencia interna de las escalas NPO y SWLS. Específicamente, se observa una diferencia sustantiva y estadísticamente significativa para la relación entre casi todas las estimaciones puntuales de las condiciones experimentales analizadas. La única excepción se encuentra en la relación entre las condiciones experimentales SWLS-3op y SWLS-5op, donde los IC y las estimaciones puntuales de α se solapan entre sí.

5.1.3 Tercera etapa: Análisis de la validez

A continuación se ofrecen los resultados observados respecto de la *evidencia basada en el grado de relación con otras variables*, a partir del modelo SEM

restringido propuesto en la figura 2 (subapartado 4.4.1, *tercera etapa*). Estos aparecen en la tabla 10. De acuerdo a dicha tabla, para la escala **NPO** se observa que X^2= 76.87 (22), $p < 0.01$, lo cual indica que el modelo restringido propuesto en la figura 2 no puede ser aceptado. Así, es necesario rechazar la hipótesis nula respecto de que la correlación entre las puntuaciones directas estimadas (*T*) con las variables criterio es igual para todas las condiciones experimentales (*dos*, *tres* y *cinco* opciones de respuesta. Por otro lado, se observa un valor RMSEA de .058, cuyo intervalo de confianza (IC) estimado fluctúa entre .03 y 0.08. Este IC indica que no se puede rechazar, al menos con un nivel de significación del 1%, la hipótesis respecto de que el modelo propuesto es una aproximación adecuada o aceptable de los datos, de acuerdo a los criterios expuestos en el capítulo 2 (Browne y Cudeck, 1993; Hair y *col.*, 1999). Lo anterior se ve reforzado por los índices CFI y TLI observados, los cuales también aparecen adecuados (CFI = .99 y TLI = .96), de acuerdo a los criterios expuestos en el capítulo antes señalado (Browne y Cudeck, 1993; Hair y *col.*, 1999).

Tabla 10: Bondad de ajuste para los modelos utilizados para analizar la validez relacionada con otras variables, en función de los distintos formatos de respuesta, para las escalas NPO y SWLS, bajo el modelo de Teoría Clásica de los Tests (TCT).

	TCT	
	NPO	SWLS
	Modelo restringido	Modelo restringido
X^2	76.87	38.62
Df	22	22
p	<.01	.02
RMSEA	.058	.042
	(.03 ; .08)	(.00 ; .07)
CFI	.99	.99
TLI	.96	.97

Nota: X^2: Chi cuadrado; *Df*: grados de libertad; *p*: p-valor; *RMSEA* = Root Mean Square Error of Approximation Estimate, con un intervalo de confianza al 99% entre paréntesis; *CFI*: Comparative Fit Index; *TLI*: Turker-Lewis Index.

Tabla 11: Evidencia basada en la relación con otras variables para los diferentes formatos de respuesta de las escalas NPO y SWLS, bajo el modelo de Teoría Clásica de los Tests (TCT).

Variables criterio	*Orientación Negativa hacia los Problemas* (NPO)			*Escala de Satisfacción con la Vida* (SWLS)		
	ρ			ρ		
	2 opciones	3 opciones	5 opciones	2 opciones	3 opciones	5 opciones
N	.75** (.71 ; .78)	.74** (.70 ; .78)	.74** (.70 ; .78)	-.40** (-.51 ; -.29)	-.43** (-.53 ; -.32)	-.44** (-.54 ; -.34)
E	-.32** (-.41 ; -.24)	-.35** (-.44 ; -.26)	-.36** (-.45 ; -.27)	.35** (.24 ; .45)	.31** (.20 ; .42)	.36** (.26 ; .47)
O	-.13** (-.23 ; -.03)	-.06 (-.16 ; .04)	-.10* (-.21 ; .00)	-.10* (-.23 ; .02)	-.05 (-.17 ; .09)	-.04 (-.18 ; .09)
A	-.04 (-.13 ; .06)	-.07 (-.16 ; .03)	-.07 (-.17 ; .03)	.11 (-.03 ; .24)	.04 (-.09 ; .17)	.09 (-.04 ; .22)
C	-.32** (-.41 ; -.24)	-.35** (-.44 ; -.26)	-.41** (-.50 ; -.33)	.27** (.16 ; .38)	.28** (.17 ; .39)	.29** (.18 ; .41)
PA	-.39** (-.47 ; -.31)	-.38** (-.47 ; -.29)	-.40** (-.49 ; -.32)	.24** (.12 ; .36)	.28** (.16 ; .40)	.31** (.20 ; .42)
NA	.49** (.41 ; .57)	.50** (.41 ; .58)	.52** (.45 ; .60)	-.35** (-.46 ; -.23)	-.35** (-.46 ; -.24)	-.34** (-.45 ; -.22)
PPO [a]	-.49** (-.57 ; -.42)	-.48** (-.56 ; -.40)	-.52** (-.60 ; -.44)	.28** (.16 ; .40)	.37** (.25 ; .48)	.39** (.28 ; .50)
RPS [a]	-.15** (-.25 ; -.06)	-.13** (-.23 ; -.03)	-.18** (-.29 ; -.07)	.17** (.03 ; .31)	.22** (.08 ; .36)	.27** (.13 ; .42)
ICS [a]	.17** (.07 ; .26)	.19** (.10 ; .29)	.23** (.13 ; .33)	-.09 (-.22 ; .05)	-.15** (-.28 ; -.02)	-.15** (-.28 ; -.03)
AS [a]	.53** (.46 ; .60)	.56** (.49 ; .63)	.63** (.57 ; .70)	-.24** (-.37 ; -.12)	-.27** (-.39 ; -.15)	-.29** (-.41 ; -.17)

Nota: ρ = coeficiente de correlación, con un intervalo de confianza al 99% entre paréntesis. Escalas del NEO-FFI: *N* = Neuroticismo, *E* = Extraversión, *O* = Apertura a la experiencia, *A* = Amabilidad, *C* = Consciente. Escalas NEO-FFI: *N* = Neuroticismo, *E* = Extroversión, *O* = Apertura a la experiencia, *A* = Amabilidad, *C* = Consciente; Escalas PANAS: *PA* = Afectos Positivos, *NA* = Afectos Negativos; Escalas SPSI-R [formas Larga (L) y Corta (S)]: *PPO* = Orientación Positiva hacia los Problemas, *RPS* = Estilo Racional de Solución de Problemas, *ICS*= Estilo Impulsivo/Descuidado, *AS*= Estilo Evitativo.

* $p < .05$; ** $p < .01$.

[a] La escala NPO fue relacionada con el SPSI-RL (forma larga), mientras que la escala SWLS fue relacionada con el SPSI-RS (forma corta).

En el entendido de que se rechaza el modelo restringido propuesto en la figura 2, lo que sigue corresponde al examen de las correlaciones observadas

entre puntuaciones directas estimadas (*T*) con las variables criterio, para cada condición experimental, a partir del modelo libre (si restricciones). Así, la tabla 11 muestra estas correlaciones y sus respectivos intervalos de confianza (IC), estimados al 99%. Como puede observarse en la tabla antes señalada, los IC para la correlación entre las distintas formas experimentales de la escala NPO y las variables criterio se solapan entre sí en la mayor parte de los casos.

La única excepción se encuentra en la relación entre NPO-2op y Estilo Evitativo (AS) del SPSI-R y en la relación entre NPO-5op y AS. En otras palabras, la correlación entre las distintas condiciones experimentales de la escala NPO y las variables externas parecen indicar que el número de opciones de respuesta utilizado parece afectar de forma poco sustantiva la *evidencia basada en el grado de relación con otras variables* de la escala NPO.

En cuanto a la escala **SWLS**, los índices de bondad de ajuste para el modelo SEM restringido propuesto en la figura 2 (subapartado 4.4.1, *tercera etapa*) aparecen en la tabla 10. Como puede observarse en dicha tabla, las estimaciones sugieren los siguientes valores: X^2= 38.62 (22), p 0.02. Estos resultados indican que el modelo restringido propuesto en la figura 2 no puede ser aceptado. Así, al igual que para el caso de la escala NPO, es necesario rechazar la hipótesis nula respecto de que la correlación entre las puntuaciones directas estimadas (*T*) con las variables criterio es igual para todas las condiciones experimentales (*dos*, *tres* y *cinco* opciones de respuesta). Por su parte, el índice RMSEA muestra un valor de .042, con un IC estimado para el mismo que fluctúa entre .00 y .07. Esto sugiere estar frente a una representación adecuada o aceptable de los datos, de acuerdo a los criterios expuestos en el capítulo 2 (Browne y Cudeck, 1993; Hair y *col.*, 1999). Paralelamente, los índices TLI y CFI parecen reforzar esta afirmación, en tanto vuelven a sugerir que el modelo propuesto puede ser considerado una representación adecuada de los datos (CFI = .99 y TLI = .97), en función

de los criterios expuestos en el capítulo antes señalado (Browne y Cudeck, 1993; Hair y *col.*, 1999).

Por su parte, la tabla 11 muestra las correlaciones observadas entre puntuaciones directas estimadas (*T*) con las variables criterio, para cada condición experimental, a partir del modelo libre (si restricciones). Al igual que para la escala NPO, se procede a examinar los IC al 99% para cada una de las correlaciones estimadas. Dicho examen, según se aprecia en la tabla 12, muestra que los IC para todas las correlaciones entre las condiciones experimentales y las distintas variables criterio se solapan entre sí. Por lo tanto, estos resultados sugieren que el número de opciones de respuesta utilizado no afecta necesariamente la *evidencia basada en el grado de relación con otras variables* de la escala SWLS.

En síntesis, los resultados observados sugieren que la *evidencia basada en el grado de relación con otras variables*, tanto para la escala NPO como para la escala SWLS, se ve afectada de forma leve o moderada por el número de opciones de respuesta. Cabe destacar que el examen de los distintos IC estimados, los cuales aparecen en la tabla 11, sugiere que existen diferencias estadísticamente significativas solo en 2 de los distintos casos analizados (para la relación entre NPO-2op y Estilo AS y para la relación entre NPO-5op y AS). Así, es posible concluir que, aunque los resultados analizados sugieren que el número de opciones de respuesta afecta de forma leve o moderada la *evidencia basada en el grado de relación con otras variables* de las escalas NPO y SWLS, éste efecto puede ser considerado poco sustantivo.

5.2 Análisis Factorial de los Ítems

En esta sección se analizan los resultados observados para las escalas NPO y SWLS, de acuerdo a los criterios expuestos en el subapartado 4.4.2. Así, en primer lugar se presentan los *análisis preliminares* (subapartado 5.2.1), luego el *análisis de la fiabilidad* (subapartado 5.2.2) y, finalmente, el *análisis de la validez* (subapartado 5.2.3), desde la perspectiva del modelo de Análisis Factorial de los Ítems (AFI).

5.2.1 Primera etapa: Análisis preliminares

La tabla 12 muestra la correlación entre los *factores* (η) de las distintas condiciones experimentales (dos, tres y cinco opciones de respuesta), para las escalas NPO y SWLS, de acuerdo al modelo ilustrado en la figura 3.

Tabla 12: Correlación entre los diferentes tipos de formatos de respuesta para las escalas NPO y SWLS, desde el modelo de Análisis Factorial de los Ítems (AFI).

Formatos de respuesta relacionados	AFI			
	NPO		SWLS	
	ρ	*IC*	ρ	*IC*
2 y 3	.94**	(.91 ; .97)	.94**	(.88 ; .99)
2 y 5	.92**	(.90 ; .94)	.93**	(.88 ; .98)
3 y 5	.93**	(.92 ; .95)	.95**	(.92 ; .99)

Nota: ρ = Coeficiente de correlación; *IC* = Intervalo de confianza al 99 % para ρ. AFI = Análisis Factorial de los Ítems. NPO = Orientación Negativa hacia los Problemas; SWLS = Satisfacción con la Vida.
** p < .01.

De acuerdo a la tabla 12, para la escala **NPO** se observa que la correlación entre los factores (η) de las diferentes condiciones experimentales tiene un rango que varía entre .92 y .94, mientras que ninguno de los intervalos de confianza al 99% incluye una correlación probable de uno. En tal sentido, estos resulta-

dos parecen indicar que el número de opciones de respuesta puede afectar las propiedades psicométricas de la escala NPO. En todo caso, también es cierto que en consideración a que el grado de correlación entre las distintas condiciones experimentales es tan elevado, la influencia del número de opciones de respuesta sobre la fiabilidad y validez de la escala NPO puede ser mínimo.

En cuanto a la escala **SWLS**, se observan resultados similares a los señalados para la escala NPO. Estos aparecen en la tabla 12, a partir de la cual se observa que el rango de correlación entre los factores (η) de las distintas condiciones experimentales fluctúa entre .93 y .95. Asimismo, ninguno de los IC al 99% sugiere una correlación probable de uno. A pesar de ello, el límite superior de estos es cercano a uno para todas las condiciones analizadas. Por lo tanto, al igual que en el caso de la escala NPO, se espera que la influencia de variar el número de opciones de respuesta tenga poca influencia sobre las propiedades psicométricas de la escala SWLS.

En su conjunto, los resultados observados en la tabla 12 sugieren que, aunque escaso, es probable apreciar mayor influencia del número de opciones de respuesta sobre la fiabilidad y la validez de la escala NPO que sobre la escala SWLS. Asimismo, los resultados analizados sugieren que el efecto del número de opciones de respuesta sobre sus propiedades psicométricas de las escalas NPO y SWLS puede ser mínimo.

5.2.2 Segunda etapa: Análisis de la fiabilidad

De acuerdo a lo señalado en el subapartado 4.4.2, a continuación se ofrecen los resultados observados para la consistencia interna, evaluado a través del coeficiente omega (ω), de las escalas NPO y SWLS. La tabla 13 muestra las esti-

maciones de ω para las escalas NPO y SWLS, bajo los supuestos del modelo de Análisis factorial de los Ítems (AFI).

Tabla 13: Consistencia Interna para las escalas NPO y SWLS, bajo los supuestos del modelo de Análisis factorial de los Ítems (AFI).

Número de opciones de respuesta	ω	
	NPO	SWLS
2	.78 (.71 ; .86)	.62 (.55 ; .69)
3	.84 (.80 ; .87)	.71 (.66 ; .77)
5	.88 (.86 ; .90)	.76 (.71 ; .81)

Nota: ω = Coeficiente Omega. Entre paréntesis se incluyen los intervalos de confianza al 99% para ω.

Según puede observarse en la tabla 13, las estimaciones de ω para las escalas NPO y SWLS son idénticos a los reportadas anteriormente para los coeficientes alfa (α) de ambas escalas, dentro del modelo TCT. Esto era esperado, por cuanto es bien sabido que cuando se calcula α y ω desde una misma muestra de datos, tales estimaciones suelen ser invariablemente similares (McDonald, 1999; Muñiz, 2003). En cuanto a los intervalos de confianza al 99% reportados para ω, estos son mayores a los observados para α. Lo anterior se debe al hecho que desde el modelo AFI es necesario calcular primero las cargas factoriales y unicidades, por lo que se cuenta con mayor cantidad de información al momento de estimar los intervalos de confianza para ω, respecto de los intervalos de confianza para α (McDonald, 1999).

Dicho lo anterior, los resultados que aparecen en la tabla 13 indican que, tanto para la escala NPO como para la escala SWLS, el valor de la estimación

puntual de ω incrementa a medida que aumenta el número de opciones de respuesta. Para la escala **NPO**, desde .78 (NPO-2op) hasta .88 (NPO-5op), y para la escala **SWLS**, desde .62 (SWLS-2op) hasta .76 (SWLS-5op). El examen de los intervalos de confianza (IC) estimados sugiere que estos se encuentran solapados entre casi todas las condiciones experimentales analizadas, salvo para la relación entre las condiciones SWLS-2op y SWLS-5op. Por su parte, tanto para la escala NPO como para la escala SWLS, la mayor parte de las estimaciones puntuales de ω caen fuera de los límites del IC del resto de condiciones experimentales, salvo para la relación entre las condiciones SWLS-3op y SWLS-5op. Así, es posible concluir que el número de opciones de respuesta afecta de forma leve o moderada la fiabilidad de las escalas analizadas. Específicamente, en función de las condiciones experimentales analizadas, se observa una diferencia sustantiva y estadísticamente significativa para la relación de casi todas las estimaciones puntuales de ω. Esto se aplica tanto a la escala NPO como a la escala SWLS.

5.2.3 Tercera etapa: Análisis de la validez

En esta sección se analizan, como se indicó en el subapartado 4.4.2, los resultados observados para (a) la evidencia basada en la relación con otras variables (subapartado 5.2.3.1) y (b) la evidencia basada en la estructura interna (subapartado 5.2.3.2) de las escalas NPO y SWLS.

5.2.3.1. Evidencia basada en la relación con otras variables

Respecto de la validez *basada en la relación con otras variables*, desde la perspectiva del modelo de Análisis Factorial de los Ítems (AFI), la tabla 14 muestra los índices de bondad de ajuste para el modelo restringido (ver figura

4) y no restringido, tanto de la escala NPO como de la escala SWLS, de acuerdo a los criterios especificados en el subapartado 4.4.2.

Tabla 14: Bondad de ajuste para los modelos utilizados para analizar la validez relacionada con otras variables, en función de los distintos formatos de respuesta, para las escalas NPO y SWLS, bajo el modelo de Análisis Factorial de los ítems (AFI).

	NPO		*SWLS*	
	No restringido	*Restringido*	*No restringido*	*Restringido*
X^2	1574.84	1637.75	279.12	312.37
Df	669	691	204	226
p	<.01	<.01	<.01	<.01
RMSEA	.04	.04	.03	.03
	(.04 ; .05)	(.04 ; .05)	(.01 ; .04)	(.02 ; .04)
CFI	.94	.94	.99	.98
TLI	.93	.93	.98	.98

Nota: X^2: Chi cuadrado; *Df*: grados de libertad; *p*: p-valor; *RMSEA* = Root Mean Square Error of Approximation Estimate, con un intervalo de confianza al 99% entre paréntesis; *CFI:* Comparative Fit Index; *TLI*: Turker-Lewis Index.

En primer lugar, mencionar que los índices de bondad ajuste *aproximada* que aparecen en la tabla 16 (véase índices RMSEA, CFI y TLI) indican que los modelos restringido y libre (no restringido) propuestos para las escalas NPO y SWLS muestran un ajuste adecuado. Así, la hipótesis de bondad de ajuste aproximada para ambos modelos (libre y restringido) no puede ser rechazada. Sin embargo, la hipótesis de bondad de ajuste *absoluta*, respecto de que los modelos analizados reproducen de forma exacta los datos, debe ser rechazada. Esto como consecuencia de que el *p* valor asociado a X^2 es inferior a .01, tanto para el modelo libre como para el modelo restringido propuesto para evaluar la bondad de ajuste de las escalas NPO y SWLS.

Respecto de la comparación de los modelos restringido y libre (no restringido), tanto para la escala NPO como para la escala SWLS, se utiliza el test

descrito por B.O. Muthén y Muthén (2005)[16] para comparar modelos anidados, de acuerdo a los criterios expuestos en el subapartado 4.4.2 (véase *tercera etapa*). Así, para la escala NPO se obtiene una estimación $X^2 = 60.95$ (22) y un $p < 0.01$. Esto indica que la correlación entre las distintas condiciones experimentales y las variables externas presenta diferencias significativas. Por otro lado, para la escala SWLS, la hipótesis acerca de que la correlación entre las distintas condiciones experimentales y las variables externas es igual, no puede ser rechazada en un nivel de significación del cinco por ciento: $X^2 = 32.37$ (22) y un $p < 0.07$.

En otras palabras, la comparación hecha entre los modelos libre y restringido, tanto para la escala NPO como para la escala SWLS, permite sugerir que es muy probable que: (a) se aprecie algún grado de efecto del número de opciones de respuesta sobre la evidencia *basada en la relación con otras variables* para la escala NPO y (b) *no* se observe efecto del número de opciones de respuesta sobre la evidencia *basada en la relación con otras variables* para la escala SWLS.

La tabla 15 presenta la correlación entre los factores (η) de las distintas condiciones experimentales (*dos*, *tres* y *cinco* opciones de respuesta) y las variables criterio (N, E, O, A, C, PA, NA, PPO, RPS, ICS y AS), tanto para la escala NPO como para la escala SWLS.

Como se observa en la tabla 15, al ser analizados los intervalos de confianza (IC) al 99% para la correlación entre los η de las diferentes condiciones experimentales de la escala NPO y las variables criterio, se observan tan solo dos diferencias estadísticamente significativas. Éstas son para la relación entre las condiciones experimentales NPO-2op y NPO-5op con la escala Neuroticis-

[16] Este test está basado, a su vez, en la prueba *Chi cuadrado escalado de Satorra-Bentler* (Satorra-Bentler Scaled Chi-Square), sin supuestos de normalidad, descrita por Satorra y Bentler (2001).

mo (N) del NEO-FFI, y para la relación entre NPO-2op y NPO-5op con la escala Consciente (C) del NEO-FFI. Para el resto de las condiciones experimentales analizadas no se observan diferencias estadísticamente significativas

Tabla 15: relación entre las distintas condiciones experimentales y las variables externas, para las escalas NPO y SWLS, bajo los supuestos del modelo de Análisis Factorial de los Ítems (AFI).

Variables Criterio	*Orientación Negativa hacia los Problemas* (NPO)			*Escala de Satisfacción con la Vida* (SWLS)		
	ρ			ρ		
	2 opciones	3 opciones	5 opciones	2 opciones	3 opciones	5 opciones
N	.84** (.81 ; .88)	.82** (.78 ; .85)	.79** (.75 ; .83)	-.53** (-.65 ; -.41)	-.53** (-.63 ; -.42)	-.53** (-.63 ; -.43)
E	-.36** (-.45 ; -.26)	-.37** (-.47 ; -.28)	-.37** (-.46 ; -.28)	.45** (.32 ; .58)	.40** (.27 ; .52)	.44** (.33 ; .56)
O	-.13** (-.23 ; -.02)	-.05 (-.16 ; .05)	-.09* (-.19 ; -.02)	-.13* (-.27 ; .02)	-.05 (-.19 ; .10)	-.05 (-.19 ; .09)
A	-.03 (-.13 ; .08)	-.06 (-.16 ; .04)	-.05 (-.16 ; .05)	.16* (-.00 ; .32)	.08 (-.07 ; .23)	.13* (-.02 ; .27)
C	-.37** (-.46 ; -.28)	-.39** (-.48 ; -.30)	-.44** (-.52 ; -.35)	.33** (.19 ; .47)	.31** (.18 ; .44)	.32** (.19 ; .45)
PA	-.44** (-.52 ; -.35)	-.41** (-.51 ; -.32)	-.42** (-.52 ; -.33)	.32** (.18 ; .46)	.34** (.21 ; .47)	.37** (.24 ; .49)
NA	.57** (.48 ; .65)	.55** (.46 ; .64)	.56** (.48 ; .64)	-.46** (-.60 ; -.32)	-.43** (-.55 ; -.30)	-.40** (-.53 ; -.28)
PPO [a]	-.54** (-.62 ; -.47)	-.52** (-.60 ; -.44)	-.54** (-.62 ; -.46)	.38** (.24 ; .52)	.45** (.32 ; .57)	.46** (.33 ; .58)
RPS [a]	-.17** (-.27 ; -.06)	-.14** (-.25 ; -.03)	-.19** (-.30 ; -.08)	.21** (.04 ; .37)	.26** (.11 ; .42)	.31** (.16 ; .47)
ICS [a]	.17** (.07 ; .28)	.20** (.09 ; .31)	.23** (.12 ; .33)	-.09 (-.25 ; .06)	-.17** (-32 ; -.02)	-18** (-.32 ; -.04)
AS [a]	.57** (.50 ; .64)	.59** (.52 ; .66)	.63** (.57 ; .70)	-.32** (-.47 ; -.17)	-.33** (-.47 ; -.19)	-.34** (-.49 ; -.20)

Nota: ρ = coeficiente de correlación, con un intervalo de confianza al 99% entre paréntesis. Escalas NEO-FFI: *N* = Neuroticismo, *E* = Extroversión, *O* = Apertura a la experiencia, *A* = Amabilidad, *C* = Consciente; Escalas PANAS: *PA* = Afectos Positivos, *NA* = Afectos Negativos; Escalas SPSI-R [formas Larga (L) y Corta (S)]: *PPO* = Orientación Positiva hacia los Problemas, *RPS* = Estilo Racional de Solución de Problemas, *ICS*= Estilo Impulsivo/Descuidado, *AS*= Estilo Evitativo.
* $p < .05$; ** $p < .01$.
[a] La escala NPO fue relacionada con el SPSI-RL (forma larga), mientras que la escala SWLS fue relacionada con el SPSI-RS (forma corta).

Por su parte, la tabla 15 sugiere que no se observan diferencias estadísticamente significativas para la correlación entre los η de las diferentes condiciones experimentales de la escala SWLS y las variables criterio, en tanto no se aprecian IC solapados entre sí.

En su conjunto, los resultados observados en este subapartado sugieren que no existe una influencia del número de opciones de respuesta sobre la validez *basada en la relación con otras variables* de la escala SWLS. En cuanto a la escala NPO, aunque se aprecian algunas diferencias estadísticamente significativas para la correlación entre algunas condiciones experimentales y ciertas variables criterio, estas pueden considerarse poco sustantivas. Así, es posible concluir que el número de opciones de respuesta afecta de forma leve, aunque no sustantivamente, la evidencia basada en al relación con otras variables de las escalas NPO y SWLS.

5.2.3.2. Evidencia basada en la estructura interna

Respecto de la ***validez estructural*** *(evidencia basada en la estructura interna*), la tabla 16 muestra los índices de bondad de ajuste de los modelos unifactoriales, propuestos en el subapartado 4.4.2 (ver figura 5), para las escalas NPO y SWLS.

Tabla 16: Evidencia basada en la estructura interna para los diferentes tipos de formatos de respuesta, para la escalas NPO y SWLS, bajo el modelo de Análisis Factorial de los Ítems (AFI).

	NPO			SWLS		
	2 opciones	3 opciones	5 opciones	2 opciones	3 opciones	5 opciones
X2	185.09	223.05	408.08	13.29	12.95	17.25
Df	35	35	35	5	5	5
p	<.01	<.01	<.01	.02	.02	.00
RMSEA	.08	.09	.12	.06	.06	.08
	(.06 ; .09)	(.07 ; .10)	(.10 ; .14)	(.00 ; .12)	(.00 ; .12)	(.01 ; .14)
CFI	.89	.90	.86	.96	.98	.97
TLI	.86	.88	.83	.93	.96	.95

Nota: X^2: Chi cuadrado; *Df*: grados de libertad; *p*: p-valor; *RMSEA* = Root Mean Square Error of Approximation Estimate, con un intervalo de confianza al 99% entre paréntesis; *CFI:* Comparative Fit Index; *TLI*: Turker-Lewis Index.

Según se observa en la tabla 16, los modelos unifactoriales propuestos para cada una de las condiciones experimentales de la escala **NPO** muestran una bondad de ajuste *absoluta* inadecuada en todos los casos analizados, en tanto el índice chi-cuadrado (X^2) estimado para cada condición arroja un *p* valor < .01. Asimismo, los índices de bondad de ajuste aproximada (véase RMSEA, CFI y TLI) parecen apuntar en la misma dirección. Así, para RMSEA se observa que el modelo NPO-2op ajusta de forma marginalmente aceptable (RMSEA = .08), mientras que para las condiciones experimentales restantes no se observa ajuste alguno (RMSEA NPO-3op = .09 y RMSEA NPO-5op = .12). Algo similar ocurre con los índices CFI y TLI, donde tampoco se observa ajuste, salvo para el modelo NPO 3 opciones de respuesta, donde el índice CFI muestra un ajuste marginalmente aceptable (CFI NPO-3op= .90).

A pesar de la falta de bondad de ajuste absoluto y/o aproximado observado para los modelos unifactoriales de las distintas condiciones experimentales de la escala NPO estimados, es posible inferir a partir de los resultados en la tabla 18 que a medida que aumenta el número de opciones de respuesta, la estructura interna (bondad de ajuste) de la escala NPO va perdiendo sistemáticamente fuerza, llegando incluso a perder todo tipo de ajuste cuando se analiza bajo la condición experimental con cinco opciones de respuesta (NPO-5op).

Por su parte, la tabla 16 sugiere un patrón de resultados similar para la valoración de la consistencia interna de la escala **SWLS**, respecto de lo señalado para la escala NPO. Según se observa en la dicha tabla, a medida que aumenta el número de opciones de respuesta, la bondad de ajuste de la escala SWLS pierde sistemáticamente fuerza. De todos modos, a diferencia de NPO, la escala SWLS muestra un ajuste X^2 adecuado para las condiciones con dos (SWLS-2op) y tres (SWLS-3op) opciones de respuesta (p < .02), mientras que el modelo unifactorial con cinco alternativas de respuesta (SWLS-5op) es rechazado en un ni-

vel de significación del 1%. Sobre este punto, cabe destacar que la estimación X^2 para la condición con tres opciones de respuesta (X^2= 12.95 (5), p < .02) se ajusta mejor que las condiciones con dos (X^2= 13.29 (5), p < .02) y cinco (X^2= 17.25 (5), p < .00) alternativas de respuesta. Así, el modelo SWLS-3op muestra las mejores estimaciones de X^2, seguido por el modelo SWLS-2op, en tanto que la condición experimental SWLS-5op aparece como la menos adecuada.

En cuanto a los índices de bondad de ajuste aproximada para las distintas condiciones experimentales de la escala SWLS, RMSEA muestra un ajuste aceptable para las condiciones SWLS-2op y SWLS-3op (RMSEA = .06, en ambos casos), en tanto uno marginalmente aceptable para la condición SWLS-5op (RMSEA = .08). A su vez, los índices CFI y TLI se observan aceptables para todos los modelos unifactoriales analizados.

En términos generales, los resultados observados en la tabla 16 sugieren que el número de opciones de respuesta tiene un efecto sustantivo sobre la consistencia de la estructura interna (bondad de ajuste) de las escalas NPO y SWLS. Así, a medida que aumenta el número de alternativas de respuesta, la estructura interna (bondad de ajuste) de las escalas analizadas va perdiendo ajuste, siendo esto especialmente relevante y extremo en el caso de la escala NPO (ver p. ej., NPO-5op).

5.3 Teoría de Respuesta a los Ítems

En este subapartado se presentan los resultados observados para las escalas NPO y SWLS, de acuerdo a los criterios expuestos en la sección 4.4.3. Así, en primer lugar se presentan los *análisis preliminares* (subapartado 5.3.1), luego el *análisis de la fiabilidad* (subapartado 5.3.2) y, finalmente, el *análisis de la validez* (subapartado 5.3.3), desde la óptica del modelo de Teoría de Respuesta a los Ítems (TRI).

5.3.1 Primera etapa: Análisis preliminares

La tabla 17 muestra la relación entre los *rasgos latentes* (θ) de las distintas condiciones experimentales (*dos*, *tres* y *cinco* opciones de respuesta), para las escalas NPO y SWLS, de acuerdo al modelo ilustrado en la figura 3.

De acuerdo a la tabla 17, la correlación entre las distintas condiciones experimentales de la escala **NPO** muestra un rango que varía entre .91 (NPO-2op) y .94 (NPO-5op). Ninguno de los intervalos de confianza (IC), estimados al 99%, muestra una correlación probable de uno en su límite superior. En todo caso, los IC para todas las relaciones entre condiciones experimentales analizadas se solapan entre si, salvo para la relación entre las condiciones NPO-2op y NPO-5op. En su conjunto, estos resultados parecen sugerir que variar el número de opciones de respuesta parece no afectar sustantivamente las propiedades psicométricas de la escala NPO.

Tabla 17: Correlación entre los diferentes tipos de formatos de respuesta para las escalas NPO y SWLS, desde el modelo de Teoría de Respuesta a los Ítems (TRI).

Formatos de respuesta relacionados	*TRI*			
	NPO		SWLS	
	ρ	*IC*	ρ	*IC*
2 y 3	.94**	(.91 ; .97)	.95**	(.89 ; 1.00)
2 y 5	.91**	(.88 ; .93)	.95**	(.90 ; 1.00)
3 y 5	.92**	(.90 ; .94)	.95**	(.91 ; .99)

Nota: ρ = Coeficiente de correlación; *IC* = Intervalo de confianza al 99 % para ρ. *TRI* = Teoría de Respuesta a los Ítems. NPO = *Orientación Negativa hacia los Problemas*; SWLS = *Escala de Satisfacción con la Vida*.
** $p < .01$.

Por otro lado, la tabla 17 muestra la correlación entre las distintas condiciones experimentales de la escala **SWLS**. Según se aprecia en dicha tabla, el valor estimado para estas correlaciones se mantiene constante entre las distintas condiciones experimentales analizadas (ρ= .95 en todos los casos). Por su parte, los IC para la correlación entre las condiciones SWLS-2op y SWLS-3op y para la correlación entre las condiciones SWLS-2op y SWLS-5op muestran un límite superior probable igual a uno, en tanto que para la correlación entre las condiciones SWLS-3op y SWLS-5op alcanza un límite superior probable de .99. De acuerdo a los criterios expuestos en el subapartado 4.4.2, estos valores sugieren que es poco probable observar efectos del número de opciones de respuesta sobre las propiedades psicométricas de la escala SWLS, bajo los supuestos del modelo TRI.

En síntesis, los resultados examinados en este apartado sugieren que el efecto del número de opciones de respuesta sobre las propiedades psicométricas de las escalas NPO y SWLS, bajo los supuestos del modelo TRI, puede ser mínimo o nulo. Es más, de observarse algún tipo de efecto, este debería apreciarse sólo sobre la escala NPO y no sobre la escala SWLS.

5.3.2 Segunda etapa: Análisis de la fiabilidad

La tabla 18 presenta la estimación de los distintos coeficientes omega (ω), junto a sus respectivos intervalos de confianza (IC), estimados al 99%, para las escalas NPO y SWLS.

Para la escala **NPO**, según se observa en la tabla 18, las condiciones con dos (NPO-2op) y tres (NPO-3op) alternativas de respuesta presentan igual consistencia interna (ω= .89). La condición con cinco (NPO-5op) opciones de respuesta es algo mas elevada, pero no sustantivamente diferente de las anteriores (ω= .92). En todo caso, los IC estimados sugieren que para la relación de ω entre las condiciones experimentales con dos (NPO-2op) y cinco (NPO-5op) opciones de respuesta o para la relación con tres (NPO-3op) y cinco (NPO-5op) alternativas de respuesta muestra diferencias estadísticamente significativas. De todos modos, estas diferencias pueden ser calificadas como mínimas, en el entendido de que los distintos intervalos de confianza se solapan entre sí.

Tabla 18: Consistencia Interna para las escalas NPO y SWLS, bajo el modelo de Teoría de Respuesta a los Ítems (TRI).

Numero de opciones de respuesta	ω	
	NPO	SWLS
2	.89 (.87 ; .91)	.80 (.75 ; .86)
3	.89 (.87 ; .91)	.81 (.76 ; .85)
5	.92 (.91 ; .93)	.81 (.78 ; .85)

Nota: ω = Coeficiente Omega. Entre paréntesis se incluyen los intervalos de confianza al 99% para ω.

Tal y como se indicó en los análisis preliminares, la escala **SWLS** *no* muestra diferencias sustantivas ni estadísticamente significativas para la esti-

mación de los índices ω de las distintas condiciones experimentales examinadas. Así, para las condiciones con tres (SWLS-3op) y cinco (SWLS-5op) opciones de respuesta se observa una estimación puntual de ω similar (ω= .81). Por el contrario, la condición con dos (SWLS-2op) alternativas de respuesta muestra una estimación puntual de ω levemente inferior respecto de las anteriores (ω= .80). Al mismo tiempo, los IC sugieren que las diferencias para ω, para la relación entre las distintas condiciones experimentales, *no* son estadísticamente significativas.

En su conjunto, los resultados observados sugieren que el número de opciones de respuesta afecta de forma poco sustantiva la consistencia interna de la escala NPO, en tanto que no se aprecian efectos del número de opciones de respuesta sobre la consistencia interna de la escala SWLS.

5.3.3 Tercera etapa: Análisis de la validez

En esta sección se analizan, como se indicó en el subapartado (4.4.3), los resultados observados para (a) la evidencia basada en la relación con otras variables (subapartado 5.3.3.1) y (b) la evidencia basada en la estructura interna (subapartado 5.3.3.2), de las escalas NPO y SWLS.

5.3.3.1. Evidencia basada en la relación con otras variables

Respecto de la ***validez basada en la relación con otras variables***, la tabla 19 muestra los índices de bondad de ajuste para el modelo ilustrado en la figura 4. En esta aparecen los índices de bondad de ajuste estimados para los modelos restringido y libre (no restringido) propuestos para las escalas NPO y SWLS.

Tabla 19: Bondad de ajuste para los modelos utilizados para analizar la validez relacionada con otras variables, en función de los distintos formatos de respuesta, para las escalas NPO y SWLS, bajo el modelo Teoría de Respuesta a los Ítems (TRI).

	NPO		SWLS	
	No restringido	Restringido	No restringido	Restringido
X^2	6453.34	6544.19	382.20	430
Df	669	691	204	226
p	<.01	<.01	<.01	<.01
RMSEA	.11 (.09 ; .13)	.11 (.10 ; .13)	.05 (.00 ; .11)	.05 (.01 ; .14)
CFI	.92	.95	.99	.99
TLI	.90	.94	.98	.97

Nota: X^2: Chi cuadrado; *Df*: grados de libertad; *p*: p-valor; *RMSEA* = Root Mean Square Error of Approximation Estimate, con un intervalo de confianza al 99% entre paréntesis; *CFI*: Comparative Fit Index; *TLI*: Turker-Lewis Index.

De acuerdo a la tabla 19, para la escala **NPO** los índices de bondad de ajuste absoluto observados para los modelos libre (X^2= 6453.34 (669), p < .01) y restringido (X^2= 6544.19 (691), p < .01) sugieren que ambos modelos deben ser rechazados. Por su parte, el índice de bondad de ajuste aproximada RMSEA sugiere una conclusión similar, en tanto se observa un valor asociado al mismo de .11 para ambos modelos. De todos modos, los índices CFI y TLI parecen indicar que, a pesar de la evidencia en contra señalada, los modelos libre y restringido de la escala NPO pueden ser tolerados, en tanto no muestran valores de bondad de ajuste aproximada inferiores a .90.

En cuanto a la escala **SWLS**, los índices de bondad de ajuste absoluta X^2 indican que los modelos libre (X^2= 382.20 (204), p < .01) y restringido (X^2= 430.00 (226), p < .01) deben ser rechazados. Por el contrario, el índice de bondad de ajuste aproximada RMSEA sugiere que ambos modelos pueden ser aceptados, en tanto muestran un valor asociado de .05. En todo caso, conviene tener presente que esto último debe ser tomado con precaución, en tanto los intervalos de confianza (IC) para los índices RMSEA de los modelos libre (.00 - .11) y restringido (.01 - .11), estimados al 99%, sugieren que ambos modelos podrían ser rechazados. A pesar de lo señalado, los índices CFI y TLI indican que tanto el

modelo libre como el modelo restringido muestran una bondad de ajuste aproximado adecuada, por lo que pueden ser tolerados como válidos.

En su conjunto, los resultados observados sugieren que es posible aceptar los modelos libre y restringido propuestos para las escalas NPO y SWLS. A pesar de ello, las conclusiones que surjan de su análisis deben ser tomadas con precaución, en tanto para ninguno de los modelos propuestos se observa una bondad de ajuste absoluta o aproximada del todo adecuada. Esto último no era esperado, en función de los resultados observados para el mismo análisis bajo los supuestos de los modelos TCT y AFI. Por ejemplo, desde la perspectiva del modelo AFI, los índices de bondad de ajuste aproximada (véase RMSEA, CFI y TLI) de los modelos libre y restringido propuestos para valorar la evidencia basada en la relación con otras variables de las escalas NPO y SWLS se muestran en todos los casos adecuados. De esta manera, bajo la óptica del modelo TRI, se esperaba observar resultados similares. Pensamos que esta discrepancia es consecuencia del procedimiento de análisis de datos utilizado por el modelo de respuesta graduada de Samejima (modelo TRI). Así, desde este modelo es necesario categorizar primero las escalas de respuesta experimentales en tantos umbrales como categorías de respuesta se esté utilizando (*dos*, *tres* y *cinco* opciones de respuesta, respectivamente, para cada condición experimental). Una vez estimados los umbrales, se procede a la evaluación de la bondad de ajuste de los modelos libre y restringido propuestos. Por el contrario, desde la perspectiva de los modelos TCT y AFI se trabaja con gradientes de respuesta continuos, donde no es necesario categorizar previamente las condiciones experimentales en umbrales. En definitiva, lo antes señalado implica que desde el modelo TRI utilizada, la estimación de los modelos libre y restringido propuestos para las escalas NPO y SWLS se realiza utilizando mayor cantidad de información, respecto de los modelos TCT y AFI. Es justamente este último hecho, entre otras variables que puedan estar interviniendo, lo que parece explicar la falta de ajuste absolu-

to (véase índice X^2) y aproximado (véase índice RMSEA) observado para los modelos libre y restringido propuestos desde el modelo TRI, para las escalas NPO y SWLS.

En cuanto a la comparación de los modelos libre y restringido, tanto para la escala NPO como para la escala SWLS, se utiliza el test descrito por B.O. Muthén y Muthén (2005) para comparar modelos anidados[17], de acuerdo a los criterios expuestos en el subapartado 4.4.2 (véase *tercera etapa*). Así, para la escala NPO se obtiene un índice $X^2 = 193.30$ (22) y un $p < 0.01$. Esto indica que la correlación entre las distintas condiciones experimentales y las variables externas presenta diferencias significativas. Lo mismo ocurre con la escala SWLS, donde se observa un índice $X^2 = 129.91$ (22) y un $p < 0.01$. En todo caso, lo antes señalado debe ser tomado con precaución, en tanto los análisis preliminares reportados en el subapartado 5.3.1 sugieren que el efecto del número de opciones de respuesta sobre las propiedades psicométricas de las escalas bajo estudio puede ser mínimo (especialmente para la escala NPO) o nulo (especialmente para el caso de la escala SWLS).

Siendo coherentes con lo sugerido en el párrafo anterior, se procede a analizar la correlación entre los rasgos latentes (θ) de las distintas condiciones experimentales (*dos*, *tres* y *cinco* opciones de respuesta) y las variables criterio (N, E, O, A, C, PA, NA, PPO, RPS, ICS y AS), tanto para la escala NPO como para la escala SWLS. Los resultados se presentan en la tabla 20.

[17] Este test está basado, a su vez, en la prueba *Chi cuadrado escalado de Satorra-Bentler* (Satorra-Bentler Scaled Chi-Square), sin supuestos de normalidad, descrita por Satorra y Bentler (2001).

Tabla 20: relación entre las distintas condiciones experimentales y las variables criterio, para las escalas NPO y SWLS, bajo los supuestos del modelo de Teoría de Respuesta a los Ítems (TRI).

Variable criterio	*Orientación Negativa hacia los Problemas* (NPO)			*Escala de Satisfacción con la Vida* (SWLS)		
	ρ			ρ		
	2 opciones	3 opciones	5 opciones	2 opciones	3 opciones	5 opciones
N	.84** (.80 ; .88)	.81** (.77 ; .85)	.77** (.73 ; .81)	-.53** (-.66 ; -.41)	-.51** (-.62 ; -.40)	-.51** (-.61 ; -.42)
E	-.36** (-.46 ; -.27)	-.37** (-.46 ; -.29)	-.37** (-.44 ; -.29)	.45** (.33 ; .57)	.38** (.25 ; .51)	.43** (.31 ; .54)
O	-.14** (-.25 ; -.04)	-.07 (-.17 ; .04)	-.11** (-.20 ; -.02)	-.14* (-.30 ; .02)	-.05 (-.19 ; .10)	-.04 (-.18 ; .09)
A	-.04 (-.14 ; .07)	-.07 (-.17 ; .03)	-.07 (-.16 ; .03)	.15* (-.01 ; .31)	.06 (-.07 ; .20)	.11* (-.02 ; .24)
C	-.37** (-.45 ; -.28)	-.39** (-.47 ; -.30)	-.42** (-.50 ; -.35)	.35** (.21 ; .49)	.32** (.19 ; .44)	.33** (.21 ; .45)
PA	-.44** (-.52 ; -.35)	-.42** (-.50 ; -.33)	-.42** (-.49 ; -.35)	.32** (.18 ; .46)	.34** (.20 ; .47)	.36** (.24 ; .48)
NA	.55** (.46 ; .63)	.54** (.46 ; .62)	.54** (.47 ; .60)	-.45** (-.58 ; -.32)	-.41** (-.53 ; -.28)	-.39** (-.49 ; -.28)
PPO [a]	-.55** (-.63 ; -.47)	-.52** (-.60 ; -.45)	-.54** (-.60 ; -.47)	.38** (.25 ; .51)	.44** (.32 ; .55)	.44** (.32 ; .56)
RPS [a]	-.17** (-.27 ; -.06)	-.14** (-.25 ; -.04)	-.19** (-.28 ; -.11)	.22** (.07 ; .36)	.26** (.13 ; .38)	.31** (.20 ; .43)
ICS [a]	.18** (.08 ; .28)	.20** (.11 ; .30)	.23** (.15 ; .31)	-.11 (-.26 ; .05)	-.17** (-.30 ; -.04)	-.19** (-.32 ; -.06)
AS [a]	.58** (.51 ; .64)	.60** (.53 ; .66)	.64** (.58 ; .69)	-.32** (-.46 ; -.18)	-.32** (-.44 ; -.20)	-.33** (-.44 ; -.22)

Nota: ρ = coeficiente de correlación, con un intervalo de confianza al 99% entre paréntesis. Escalas NEO-FFI: *N* = Neuroticismo, *E* = Extroversión, *O* = Apertura a la experiencia, *A* = Amabilidad, *C* = Consciente; Escalas PANAS: *PA* = Afectos Positivos, *NA* = Afectos Negativos; Escalas SPSI-R [formas Larga (L) y Corta (S)]: *PPO* = Orientación Positiva hacia los Problemas, *RPS* = Estilo Racional de Solución de Problemas, *ICS*= Estilo Impulsivo/Descuidado, *AS*= Estilo Evitativo.

* $p < .05$; ** $p < .01$.

[a] La escala NPO fue relacionada con el SPSI-RL (forma larga), mientras que la escala SWLS fue relacionada con el SPSI-RS (forma corta).

De acuerdo a los resultados observados en la tabla 20, los intervalos de confianza (IC) estimados para la correlación entre las diferentes condiciones experimentales y las variables criterio, tanto para la escala NPO como para la es-

cala SWLS, se solapan entre sí en todos los casos analizados. De esta forma y en función de los criterios expuestos en el subapartado 4.4.2, no se observa efecto del número de opciones de respuesta sobre la *validez basada en la relación con otras variables* de las escalas NPO y SWLS, desde la perspectiva del modelo TRI. Esto contradice lo señalado anteriormente, tras comparar los modelos libre y restringido para las escalas NPO y SWLS, en el sentido que se esperaba encontrar diferencias estadísticamente significativas para la correlación entre las distintas condiciones experimentales y las variables criterio. Por el contrario, estos resultados son coherentes con lo señalado en el subapartado 5.3.1, donde se señaló que el efecto del número de opciones de respuesta sobre las propiedades psicométricas de las escalas NPO y SWLS sería mínimo o nulo.

5.3.3.2. Evidencia basada en la estructura interna

Respecto de la ***validez estructural*** *(evidencia basada en la estructura interna*), la tabla 21 muestra los índices de bondad de ajuste de los modelos unifactoriales (*dos*, *tres* y *cinco* opciones de respuesta), para las escalas NPO y SWLS, propuestos en el subapartado 4.4.3 (véase *tercera etapa*, evidencia basada en la estructura interna y figura 5). Según se observa en dicha tabla, la bondad de ajuste absoluta de los modelos propuestos para la escala **NPO**, evaluados a través del índice Chi cuadrado (X^2), muestran un *p* valor asociado de .00, lo cual indica que éstos deben ser rechazados. Llama especialmente la atención la estimación de X^2 para la condición experimental con *cinco* opciones de respuesta (X^2= 1784.19 (35), p = .00), en tanto muestra valores sustancialmente extremos, respecto de los valores X^2 asociados a las condiciones experimentales con *dos* (X^2= 351.05 (35), p = .00) y *tres* (X^2= 363.23 (35), p = .00) alternativas de respuesta. Se requiere más investigaciones para explicar este fenómeno.

Tabla 21: Evidencia basada en la estructura interna para los diferentes tipos de formatos de respuesta, para la escalas NPO y SWLS, bajo el modelo de Teoría de Respuesta a los Ítems (TRI).

	NPO			SWLS		
	2 opciones	3 opciones	5 opciones	2 opciones	3 opciones	5 opciones
X^2	351.05	363.23	1784.19	9.21	17.24	24.41
Df	35	35	35	5	5	5
p	.00	.00	.00	.10	.00	.00
RMSEA	.11	.11	.26	.04	.08	.10
	(.09 ; .13)	(.10 ; .13)	(.24 ; .28)	(.00 ; .11)	(.01 ; .14)	(.04 ; .16)
CFI	.92	.95	.92	.99	.99	.99
TLI	.90	.94	.90	.98	.97	.98

Nota: X^2: Chi cuadrado; *Df*: grados de libertad; *p*: p-valor; *RMSEA* = Root Mean Square Error of Approximation Estimate, con un intervalo de confianza al 99% entre paréntesis; *CFI:* Comparative Fit Index; *TLI*: Turker-Lewis Index.

Por su parte, el índice de bondad de ajuste aproximada RMSEA, para los modelos unifactoriales propuestos para las distintas condiciones experimentales de la escala NPO (*dos*, *tres* y *cinco* opciones de respuesta), muestra conclusiones similares a las expuestas en el párrafo precedente. Así, la condición experimental con *cinco* (NPO-5op) opciones de respuestas muestra una bondad de ajuste aproximada RMSEA claramente inaceptable (RMSEA NPO-5op = .26), en tanto que las condiciones con *dos* (NPO-2op) y *tres* (NPO-3op) alternativas de respuesta presentan estimaciones que obligan a rechazar ambos modelos (RMSEA NPO-2op y NPO-3op = .11). Curiosamente, los índices de bondad de ajuste aproximada CFI y TLI apuntan en una dirección contraria, en tanto sugieren que todos los modelos unifactoriales analizados muestran un ajuste adecuado, siendo los modelos con *dos* (NPO-2op) y *cinco* (NPO-5op) opciones de respuesta los que menos ajustan (para más detalles, véase la tabla 21). Se requieren nuevas investigaciones que permitan explicar las discrepancias observadas para las conclusiones sugeridas por los distintos índices de bondad de ajuste examinados. Mientras los índices X^2 y RMSEA indican que los modelos en cuestión deben ser rechazados, los índices CFI y TLI sugieren que los modelos propuestos ajustan adecuadamente.

Por otro lado, los índices de bondad de ajuste absoluta X^2 para la escala **SWLS** indican que los modelos unifactoriales propuestos para las condiciones con *tres* (X^2= 17.24 (5), p = .00) y *cinco* (X^2= 24.41 (5), p = .00) opciones de respuesta deben ser rechazados, mientras que la condición con *dos* (X^2= 9.21 (5), p = .10) alternativas de respuesta no puede ser rechazada. Se observan resultados similares a partir de la estimación del índice de bondad de ajuste aproximada RMSEA, donde el modelo unifactorial para la condición con *dos* (SWLS-2op) categorías de respuesta vuelve a mostrar un ajuste adecuado (RMSEA SWLS-2op = .04), en tanto que las condiciones con *tres* (RMSEA SWLS-3op = .08) y *cinco* (RMSEA SWLS-5op = .10) alternativas de respuesta muestran un ajuste deficiente. Por el contrario, los índices de bondad de ajuste aproximada CFI y TLI sugieren que todos los modelos unifactoriales propuestos muestran un ajuste adecuado en todos los casos (para más detalles, véase la tabla 21). Al igual que lo señalado para la escala NPO, se requieren nuevas investigaciones que permitan explicar las discrepancias observadas para las conclusiones sugeridas por los distintos índices de bondad de ajuste examinados. Mientras los índices X^2 y RMSEA indican que los modelos en cuestión deben ser rechazados, los índices CFI y TLI sugieren que los modelos propuestos ajustan adecuadamente.

En su conjunto, los resultados analizados sugieren que el número de opciones de respuesta afecta la evidencia basada en la estructura interna de las escalas NPO y SWLS, bajo la óptica del modelo TRI. Así, se observa que el ajuste estructural de ambas escalas empeora a medida que aumenta el número de opciones de respuesta.

Sección III: Discusión y conclusiones

Llegados a este punto, la sección III presenta la discusión y conclusiones del presente trabajo de Tesis de Doctorado. Se estructura en dos capítulos. Así, el ***capítulo 6*** propone una *discusión* final acerca de los distintos tópicos considerados en este trabajo, en tanto que el ***capítulo 7*** presenta las *conclusiones* y limitaciones que derivan del mismo.

6. *Discusión*

Dentro del ámbito de la psicología científica, no cabe la menor duda que es necesario recurrir a la *medición* como medio para validar o refutar teorías, modelos o constructos psicológicos. Desconocer su necesidad y utilidad es un sin sentido y, por que no decirlo, una irresponsabilidad en tanto científicos. La sección I del presente trabajo de Tesis de Doctorado ha considerado éste tópico en profundidad, no con otro objeto que establecer un *marco teórico* de referencia al mismo. Así, nuestra adaptación del modelo de *estructura básica de las Ciencias Sociales* (ver figura 1) propuesto por Torgerson (1962) sirve de base para abordar los distintos temas tratados en este trabajo.

Por otro lado, como se indicó en la sección I, la psicología científica recurre a distintas técnicas de evaluación para *describir-explicar* un problema y/o para *valorar un tratamiento* o *intervención* efectuado (Fernández-Ballesteros, 2004a). Es más, las técnicas de evaluación son fundamentales dentro del *proceso de validación o refutación de cualquier teoría, modelo o constructo psicológico* que se quiera adoptar. Una buena praxis sugiere utilizar una combinación de estas técnicas (*observación*, *técnicas objetivas*, *autoinformes*, *entrevistas*, *técnicas proyectivas*, *técnicas psicométricas*). Entre las mismas, dada la sencillez y lo económico de su aplicación, los *cuestionarios* son considerados la herramienta más difundida (Fernández-Ballesteros, 2004b; Pelechano, 2000). Hay acuerdo entre teóricos e investigadores sobre el hecho de que la *Teoría Clásica de los Tests* (TCT), el *Análisis Factorial de los Ítems* (AFI) y la *Teoría de Respuesta a los Ítems* (TRI) son los principales modelos de referencia al momento de examinar las propiedades psicométricas de este tipo de instrumentos de evaluación (Allen y Yen, 1979; Gulliksen, 1987; Hambleton y Swaminathan, 1985; Lord, 1980; Martínez-Arias, 1996; McDonald, 1999; Muñiz, 1996; Santisteban-Requena, 1990).

Los analistas de este tipo de pruebas psicológicas señalan que la creación de todo cuestionario es fruto de un proceso que implica diversas etapas (Fernández-Ballesteros, 2004a; Muñiz, 2003, 1996; Martínez-Arias, 1996). Entre otras, se debe especificar el tipo de formato de respuesta a utilizar, el cual suele ser de tipo dicotómico, de selección múltiple o escalas tipo Likert (Muñiz, 2003, 1996; Martínez-Arias, 1996). Ahora bien, dentro del ámbito de la evaluación del *rendimiento* académico y/o profesional, por regla general, se utilizan formatos de respuesta del tipo *selección múltiple* (Fernández-Ballesteros, 2004a; Muñiz, 2003, 1996; Martínez-Arias, 1996). Para este tipo de pruebas y formato de respuesta, existe acuerdo entre teóricos e investigadores respecto de que *tres* es el número de opciones de respuesta idóneo para garantizar una adecuada fiabilidad para este tipo de instrumentos (Abad, Olea y Ponsoda, 2001; Bruno y Dirkzwager, 1995; Costin, 1970, 1972; Lord, 1977, 1980; Rogers y Harley, 1999; Straton y Catts, 1980; Tversky, 1964).

Por el contrario, dentro del terreno de la evaluación de las *actitudes* o de la *personalidad* suelen utilizarse formatos de respuesta dicotómicos y tipo Likert (Muñiz, 2003, 1996; Martínez-Arias, 1996). Entre estos, aún no queda claro cuál debe de ser el número óptimo de alternativas de respuesta (Bearden, Netmeyer y Mobley, 1993; Cox, 1980; Peter, 1979; Shaw y Wright, 1967). Las distintas investigaciones empíricas o simulaciones revisadas en este trabajo de Tesis de Doctorado sugieren conclusiones diversas y en muchos casos contradictorias entre sí. Mientras que algunos investigadores sostienen que basta con utilizar *dos* o *tres* alternativas de respuesta (Aiken, 1983; Komorita y Graham, 1965; Masters, 1974; Matell y Jacoby, 1971, 1972; Peabody, 1962; Sancerni, Meliá y González, 1990), otros sugieren usar desde *cinco* (Churchill y Peter, 1984; García-Cueto, Muñiz y Lozano, 2002) o *siete* opciones (Ferrando, 1995; Preston y Colman, 2000; Ramsay, 1973; Symonds, 1924), hasta un máximo de *25* alternati-

vas de respuesta (Champney y Marshall, 1939) para maximizar las propiedad psicométricas de este tipo de instrumentos de evaluación.

Los teóricos e investigadores discrepan también respecto de qué *método* utilizar para evaluar el efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios de personalidad. Mientras que algunos optan por recoger un número elevado de alternativas de respuestas, las cuales colapsan luego en distintos formatos de respuesta arbitrarios (p.ej., Matell y Jacoby, 1971; Peabody, 1962), otros prefieren administrar las distintas condiciones experimentales propuestas en muestras independientes, las cuales comparan entre sí (p. ej., Chang, 1994; Master, 1979; Preston y Colman, 2000). En nuestro caso, pensamos que la estrategia más adecuada consiste en permitir que cada sujeto responda una misma escala (en nuestro caso, NPO y SWLS), dentro de una misma sesión, utilizando los distintos formatos de respuesta propuestos por el investigador (p. ej., dos, tres y cinco opciones de respuesta). A nuestro juicio, esta estrategia permite recabar información directa acerca de *cómo* han contestado realmente los participantes. Así, se evitan las inferencias derivadas de utilizar los métodos antes consignados (p. ej., colapsar datos o comparara muestras distintas).

Tampoco existe consenso sobre qué *modelo psicométrico* utilizar como referente al momento de analizar el efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios de personalidad. Mientras que algunos investigadores recurren a la *Teoría Clásica de los Test* (p. ej., Aiken, 1983; Bandalos y Enders, 1996; Cox, 1980; Masters, 1974; Weng, 2004), otros sugieren utilizar el *Análisis Factorial de los Ítems* (p. ej., Comrey y Montag, 1982; Chang, 1994; Ferrando, 2000; García-Cueto *y col.*, 2002; Matell y Jacoby, 1971; Olsson, 1979) o, por el contrario, la *Teoría de Respuesta a los Ítems* (p ej., García-Cueto, Muñiz y Lozano, 2003; Hernández, Muñiz y García-Cueto, 2000).

Para examinar si el tipo de modelo psicométrico empleado tiene efectos sobre el tema en cuestión, el presente trabajo de Tesis de Doctorado realiza los distintos análisis estadísticos propuestos en la sección II utilizando simultáneamente los tres modelos psicométricos antes consignados (TCT, AFI y TRI).

En la misma línea de lo señalado hasta el momento, tampoco existe acuerdo respecto al establecimiento de un criterio unívoco que permita valorar cuál debe ser el número óptimo de opciones de respuesta. Históricamente, la mayor parte de los investigadores a optado por recurrir al examen de la *fiabilidad* como criterio, siendo el análisis del coeficiente alfa el medio más utilizado (p. ej., Aiken, 1983; Bandalos y Enders, 1996; Cicchetti, Showalter y Tyrer, 1985; Jenkins y Taber, 1977; Komorita y Graham, 1965; Lissitz y Green, 1975; Masters, 1974; Weng, 2004). Luego están las investigaciones que centran su interés en el examen de la *validez* (p. ej., Comrey y Montag, 1982; Olsson, 1979) o, preferentemente, sobre el análisis simultaneo de la fiabilidad y la validez de los cuestionarios de personalidad (p. ej., Chang, 1994; García-Cueto *y col.*, 2002; Matell y Jacoby, 1971, 1972; McCallum, Keith y Wiebe, 1988; Preston y Colman, 2000; Sancerni *y col.*, 1990). Finalmente, aunque en una medida notablemente inferior a las anteriores, es posible encontrar estudios que evalúan el efecto del número de opciones de respuesta sobre los *estadísticos descriptivos* (p. ej., Aiken, 1983; Muñiz, García-Cueto y Lozano, 2005) o sobre el *porcentaje de varianza* explicado por el factor evaluado por este tipo de cuestionarios (p. ej., Velicer y Stevensons, 1978; García-Cueto *y col.*, 2002; Muñiz *y col.*, 2005).

Por nuestra parte, pensamos que para examinar el efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios de personalidad es necesario utilizar como criterios de valoración tanto la fiabilidad (p. ej., *consistencia interna* y *estabilidad temporal*) como la validez (p. ej., *evidencia basada en la estructura interna* y *evidencia basada en la relación con otras varia-*

bles) mostrada por este tipo de cuestionarios. En cuanto al análisis de los estadísticos descriptivos, realizado históricamente bajo la óptica del modelo TCT, creemos que no aporta información sustantiva sobre el tema en cuestión. Esto como consecuencia de que a medida que aumenta el número de opciones de respuesta, debe incrementar necesariamente también la media y desviación estándar de la escala examinada. Para ilustrar este punto, considérese por ejemplo un cuestionario compuesto por ítems binarios, con una escala de respuesta en la que "0"= *En desacuerdo* y "1"= *De acuerdo* con la cuestión planteada. Si un sujeto está de acuerdo con el enunciado *X*, su puntuación en dicho enunciado debe ser 1. Por el contrario, si se utiliza un cuestionario con los mismos enunciados, pero compuesto por ítems politómicos que deben ser contestados por medio de una escala tipo Likert donde "0"= *Muy en desacuerdo*, "1"= *De desacuerdo*, "2"= *Ni de acuerdo ni en desacuerdo*, "3"= *De acuerdo* y "4"= *Muy de acuerdo* se observa lo siguiente: si el sujeto es consistente con la respuesta dada al enunciado *X* de cuestionario binario, en este caso debe contestar utilizando las etiquetas de respuesta *De acuerdo* o *Muy de acuerdo*, por lo que su puntuación sobre el enunciado *X* debe ser 3 ó 4. De observarse una puntuación de 0 ó 1, implica que el sujeto ha cambiado el sentido de su respuesta al enunciado *X* y, por tanto, está siendo inconsistente en su manera de responder. Así, este ejemplo permite ilustrar como, si los sujetos son consistentes en sus respuestas, cuando se incrementa el número de opciones de respuesta, la media y la desviación estándar de la escala debe aumentar necesariamente. Sin perjuicio de lo anterior y acorde a nuestro interés por realizar una investigación lo más exhaustiva posible, el anexo 4 describe el efecto del número de opciones de respuesta sobre los estadísticos descriptivos de las escalas NPO y SWLS, bajo los supuestos del modelo TCT. Téngase presente que lo descrito en dicho anexo debe ser considerado como una ampliación del ejemplo anterior, en tanto pensamos que no representa un aporte sustantivo al tema tratado en este trabajo de Tesis de Doctorado.

Por su parte, no se ha incluido el análisis del efecto del número de opciones de respuesta sobre el porcentaje de varianza explicado por el factor evaluado por las escalas NPO y SWLS, por cuanto estimamos que no procede realizar el mismo bajo los supuestos del modelo AFI. Examinar el porcentaje de varianza explicado por el factor evaluado por una escala *X* tiene sentido si los datos se evalúan desde el modelos de *Análisis de Componentes Principales* (ACP), no así desde el modelo de *Análisis Factorial* (AF). Aunque ambas técnicas se utilizan para reducir las variables observadas en un número más pequeño de dimensiones, el objetivo principal de todo ACP consiste en explicar la varianza total de las variables resultantes (componentes), en tanto que el objetivo fundamental de AF consiste en reproducir la matriz de correlaciones teórica a partir de los datos, en función de las comunalidades observadas (factores). Coherentes con lo señalado y sobre la base que el presente trabajo de Tesis de Doctorado utiliza como referente el modelo AFI, el capítulo 5 se ha analizado y descrito si los datos utilizados reproducen o no la matriz de correlaciones teórica de los modelos unifactoriales propuestos en el capítulo 4. Es decir, se reportan los índices de bondad de ajuste (absoluta y aproximada) y el *p*-valor asociado a los mismos respecto del grado de ajuste estructural (evidencia basada en la estructura interna) de los modelos propuestos. De todos modos, con el deseo de ser exhaustivos, en el anexo 5 se analiza el efecto del número de opciones de respuesta sobre el porcentaje de varianza explicado por el factor evaluado por las escalas NPO y SWLS, bajo la óptica del modelo ACP.

De acuerdo con la información de que disponemos, esta es la primera vez que se aborda el estudio del efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios de personalidad a través del método y procedimientos sugeridos en este trabajo de Tesis de Doctorado. A continuación se ofrece una discusión respecto del examen de los resultados analizados en la sección II del presente trabajo.

Respecto del análisis de la **bondad de ajuste** (*validez basada en la estructura interna*) de los distintos modelos unifactoriales propuestos para las escalas NPO y SWLS, en función de las condiciones experimentales utilizadas (dos, tres y cinco opciones de respuesta), se observa un claro efecto del número de opciones de respuesta sobre esta variable. Así, a medida que aumenta el número de alternativas de respuesta, la bondad de ajuste (*estructura interna*) de los instrumentos analizados pierde sistemáticamente fuerza. Esto es especialmente relevante para el caso de la escala NPO, la cual llega a presentar un ajuste francamente deficiente cuando es analizada bajo la condición experimental de cinco alternativas de respuesta (NPO 5op). En su conjunto, estos resultados son sorprendentes y no esperados, en tanto van en una dirección contraria a lo señalado por distintas investigaciones empíricas y simulaciones revisadas en este trabajo (véase apartado *estado de la cuestión*). Por ejemplo, Matell y Jacoby (1971) u Olsson (1979) sostienen el número de opciones de respuesta no afecta la estructura interna (bondad de ajuste) de la escalas de personalidad. Por su parte, nuestros resultados contradicen lo señalado por Comrey y Montag (1982), Ferrando (2000), Hernández, Muñiz y García-Cueto (2000) o García-Cueto, Muñiz y Lozano (2002), en tanto estos investigadores sugieren que a medida que aumenta el número de opciones de respuesta, incrementa también la bondad de ajuste (evidencia basada en la estructura interna) de las escalas de personalidad analizadas.

De todos modos, es necesario llevar a cabo nuevas investigaciones sobre esta cuestión, en tanto Maydeu-Olivares, Kramp y García-Forero (2006) sugieren que lo anterior puede ser consecuencia de una mala especificación del modelo que sustentan las escalas NPO y SWLS. En otras palabras, estos autores señalan que frente a un modelo incorrectamente especificado, a mayor número de opciones de respuesta tal error de especificación se hace cada vez más evidente. Por el contrario, frente a modelos correctamente especificados, no se observa

efectos del número de opciones de respuesta sobre el la bondad de ajuste del mismo (Maydeu-Olivares, Kramp y García-Forero, 2006). Por lo tanto, la pérdida sistemática de ajuste estructural (*validez basada en la estructura interna*) de las escalas NPO y SWLS analizadas en este trabajo puede deberse al hecho antes señalado y no reflejar necesariamente un efecto directo del número de opciones de respuesta sobre la bondad de ajuste de las escalas NPO y SWLS. De corroborarse lo sugerido por Maydeu-Olivares, Kramp y García-Forero (2006), ello brinda apoyo parcial a lo señalado por Komorita y Graham (1965), Matell y Jacoby (1971) u Olsson (1979), en tanto el número de opciones de respuesta parece no afectar la bondad de ajuste de modelos unifactoriales correctamente especificados.

Por su parte, el análisis de la **fiabilidad** (*consistencia interna*) de las escalas NPO y SWLS sugiere que el efecto del número de opciones de respuesta sobre esta variable está mediado por el tipo de modelo psicométrico utilizado (TCT, AFI o TRI). De acuerdo a lo señalado en el capítulo 4 (subapartado 4.4), la consistencia interna de las escalas NPO y SWLS se evaluó desde el modelo TCT por medio del coeficiente alfa (α), mientras que desde los modelos AFI y TRI se estimó a través del coeficiente omega (ω). Desde la perspectiva de los ***modelos TCT y AFI***, se observa un efecto significativo y sustantivo del número de opciones de respuesta sobre la fiabilidad (*consistencia interna*) de las escalas NPO y SWLS. Así, a medida que aumenta el número de alternativas de respuesta, incrementa sistemáticamente el valor de la estimación puntual de los coeficientes α (TCT) y ω (AFI). Por su parte, el examen de los intervalos de confianza (IC) para la estimación puntual de los distintos coeficientes α (TCT) y ω (AFI) observados, en función de cada condición experimental (dos, tres y cinco opciones de respuesta) y escala (NPO y SWLS), sugieren que este incremento es estadísticamente significativo en la mayor parte de los casos analizados. La única excepción, tanto desde el modelo TCT como desde el modelo AFI, es para la relación

entre las condiciones experimentales SWLS-3op y SWLS-5op, donde no se observan diferencias estadísticamente significativas. Finalmente, cabe destacar que las diferencias observadas para la relación entre condiciones experimentales, tanto para la escala NPO como SWLS, pueden ser calificadas de sustantivas. Por ejemplo, bajo los supuestos del modelo TCT, si se compara la estimación puntual del coeficiente α de la condición experimental NPO 2op (α= .78) con la condición experimental NPO 5op (α= .88), se observa un incremento del 13 por ciento para la estimación puntual de α entre estas condiciones experimentales. Asimismo, bajo los supuestos del modelo AFI, si se compara la estimación puntual del coeficiente ω de la condición experimental SWLS 2op (ω = .62) con la condición experimental SWLS 5op (ω = .76), se observa un incremento del 18 por ciento para la estimación puntual de ω entre estas condiciones experimentales. En su conjunto, estos resultados siguen la misma línea de lo sugerido por Symonds (1924), Pemberton (1933), Champney y Marshall (1939), Kamorita y Graham (1965), Masters (1974), Lissitz y Green (1975), Jenkins y Taber (1977), Cicchetti, Showalter y Tyrer (1985),Oaster (1989), Chang (1994), Bandalos y Enders (1996), Preston y Colman (2000), García-Cueto, Muñiz y Lozano (2002), Weng (2004) o Muñiz, García-Cueto y Lozano (2005), en tanto estos autores sugieren que el número de opciones de respuesta afecta la fiabilidad (*consistencia interna*) de las escalas de personalidad, en la dirección antes señalada. Por el contrario, los resultados reportados en este trabajo contradicen lo sugerido por Peabody (1962), Matell y Jacoby (1971), Aiken (1983), McCallum, Keith y Wiebe (1988) o Sancerini, Meliá y González-Romá (1990), puesto que estos autores señalan que el número de opciones de respuesta no afecta la fiabilidad (*consistencia interna*) de las escalas de personalidad.

Por el contrario, bajo los supuestos del ***modelo de TRI***, el número de opciones de respuesta parece no afectar la fiabilidad (*consistencia interna*) de las escalas NPO y SWLS. La estimación puntual del coeficiente ω, en función de cada

condición experimental (dos, tres y cinco opciones de respuesta) y escala (NPO y SWLS) se muestra relativamente constante. Así, para la escala NPO se observa una estimación puntual del coeficiente ω para las condiciones experimentales NPO 2op y NPO 3op de .89, en tanto que para la condición experimental NPO 5op ésta es de .92. Por su parte, para la escala SWLS se observa una estimación puntual del coeficiente ω para la condición experimental SWLS 2op de .80, mientras que para las condiciones experimentales SWLS 3op y SWLS 5op ésta es de .81. Adicionalmente, sin importar la escala analizada, el examen de los IC estimados para cada condición experimental no revela diferencias estadísticamente significativas. Por lo tanto, bajo los supuestos del modelo TRI, no se observa efecto del número de opciones de respuesta sobre la precisión (fiabilidad) de las escalas NPO y SWLS. Lo anterior brinda apoyo parcial a lo sugerido por García-Cueto, Muñiz y Lozano (2003). Según se indicó en el capítulo 3 (*Estado de la cuestión*), estos autores realizan un estudio de simulación en el que estiman la *precisión* (fiabilidad) de una escala, variando el número de opciones de respuesta desde 3 hasta 9 categorías, a partir de los *parámetros de pendiente* y la *función de información* aportada por los ítems de dicha escala. En su conjunto, los resultados ofrecidos por García-Cueto, Muñiz y Lozano (2003) indican que los parámetros de pendiente no parecen estar afectados por el número de opciones utilizado, pero si la función de información aportada por los ítems de dicha escala. Sobre esto último, la cantidad de información (precisión) ofrecida por el test aumenta sistemáticamente a medida que incrementa el número de opciones de respuesta, desde 4 hasta 9 opciones de respuesta. Paradójicamente, para la condición de 3 alternativas de repuesta se observan valores máximos, incluso superior a la cantidad de información ofrecida por el caso de 9 opciones de respuesta.

A pesar de lo señalado por García-Cueto, Muñiz y Lozano (2003), no se esperaba observar en este trabajo que el modelo psicométrico utilizado influye-

se sobre la estimación de la fiabilidad (*precisión*) de las escalas NPO y SWLS. En otras palabras, se pensaba que el número de opciones de respuesta influiría de forma sistemática, sin importar el modelo psicométrico utilizado, en las estimaciones de la consistencia interna de las escalas analizadas. Así, se requiere de nuevas investigaciones que permitan explicar la influencia del modelo psicométrico sobre el análisis de la fiabilidad (precisión) de las escalas de personalidad. Desde la perspectiva del modelo TRI se sugiere utilizar un método de estimación de la precisión de las escalas de personalidad distinto al aquí propuesto. Por ejemplo, una alternativa plausible puede ser analizar y comparar las áreas de *función de información* estimada para cada condición.

Por el contrario, si se esperaba que desde el modelo TRI las estimaciones del coeficiente ω fuesen superiores a las estimadas desde el modelo AFI o a la estimación de α, desde el modelo TCT. Esto queda reflejado en los resultados expuestos en la sección II y es consecuencia directa de la forma de estimar cada uno de estos coeficientes, en función del modelo pricométrico utilizado como referente. De acuerdo a la teoría psicométrica, para una misma muestra de datos, las estimaciones del coeficiente α, derivado desde el modelo TCT, y del coeficiente ω, derivado desde el modelo AFI, deben ser iguales para un mismo número de opciones de respuesta (McDonald, 1999; Muñiz, 2003). Asimismo, se espera que el coeficiente ω, estimado desde el modelo TRI, sea mayor que el calculado desde el modelo AFI y, por consiguiente, mayor que el coeficiente α (modelo TCT). Esto se debe a que el primero es estimado utilizando correlaciones Policóricas y los segundos por medio de correlaciones de Pearson. Invariablemente, las correlaciones Policóricas son mayores que las correlaciones de Pearson (McDonald, 1999; Olsson, 1979). Lo anterior puede ejemplificarse a partir de los resultados observados en la sección II del presente trabajo. Así, por ejemplo, la fiabilidad (precisión) para dos opciones de respuesta del modelo NPO-TRI (ω= .89) es similar a la observada para el modelo NPO-TCT (α= .88) o

NPO-AFI (ω= .88) con 5 opciones de respuesta, en tanto que la fiabilidad para dos opciones de respuesta del modelo SWLS-TRI (ω= .80) es mayor que la mostrada para el modelo SWLS-TCT (α= .76) o SWLS-AFI (ω= .76) con 5 opciones de respuesta. Así, los modelos TCT o AFI requieren más alternativas de respuesta para obtener una cantidad similar o inferior de información que la conseguida desde el modelo de TRI, con pocas opciones de respuesta.

Contrariamente a lo señalado hasta el momento, los resultados observados parecen indicar que el número de opciones de respuesta no afecta de forma sustantiva la **validez basada en la relación con otras variables**. Es decir, bajo los supuestos del modelo TCT, cuando se compara la correlación entre la puntuación total o directa (T) de cada condición experimental con sus respectivas variables criterio (N, E, O, A, C, PA, NA, PPO, RPS, ICS y AS), no se observa un efecto sustantivo del número de opciones de respuesta sobre esta relación. Esto se aplica tanto a la escala NPO como a la escala SWLS. Por su parte, se observan resultados similares desde la perspectiva de los modelos AFI y TRI. De esta forma, cuando se compara la correlación entre el factor (η) (modelo AFI) o el rasgo latente (θ) (modelo TRI) estimado para cada condición experimental con sus respectivas variables criterio (N, E, O, A, C, PA, NA, PPO, RPS, ICS y AS), no se observan efectos sustantivos del número de opciones de respuesta sobre esta relación, ya sea para la escala NPO como para la escala SWLS. Lo anterior se justifica a partir del análisis de los distintos intervalos de confianza (IC) estimados, en función de cada modelo psicométrico (TCT, AFI y TRI) y escala (NPO y SWLS), para la correlación entre las diferentes condiciones experimentales y las variables criterio propuestas. Así, el análisis de los distintos IC estimados sugiere que solo se observan casos aislados de diferencias estadísticamente significativas para las correlaciones antes señaladas[18]. En su conjunto, és-

[18] Específicamente, bajo los supuestos del modelo TCT, se observan diferencias estadísticamente significativas sólo para las relaciones entre NPO-2op y NPO-5op con AS del SPSI-R. Por el con-

tas son poco sustantivas e incluso pueden ser calificadas de irrelevantes si se considera el elevado número de variables y condiciones experimentales analizadas. Por lo tanto, es posible concluir que el número de opciones de respuesta no afecta de forma sustantiva la evidencia basada en la relación con otras variables. Estos resultados brindan apoyo a lo sugerido por Matell y Jacoby (1971), Chang (1994) o Preston y Colman (2000), en tanto éstos investigadores sugieren conclusiones similares a las reportadas en este trabajo. Por el contrario, los resultados observados parecen contradecir lo señalado por Sancerini, Meliá y González-Romá (1990), en el entendido que estos autores indican que a medida que incrementa el número de opciones de respuesta, aumenta la validez de criterio (*validez basada en la relación con otras variables*) de las escalas de personalidad por ellos analizadas.

En lo que se refiere a la **fiabilidad test-retest** (*estabilidad temporal*), tampoco se observa influencia del número de opciones de respuesta sobre ésta variable. Bajo los supuestos del modelo TCT, la estimación puntual de la correlación test-retest para las distintas condiciones experimentales de la escala SWLS incrementa a medida que aumenta el número de opciones de respuesta (SWLS 2op= .88; SWLS 3op= .89; SWLS 5op= .91). A pesar de ello, el examen de los distintos IC estimados para dicha correlación sugieren que este incremento no es estadísticamente significativo, en tanto todos los IC se encuentran solapados entre sí. De esta forma, los resultados observados indican que el número de opciones de respuesta parece no afectar la fiabilidad test-retest de la escala SWLS, lo cual es consistente con lo sugerido por Matell y Jacoby (1971) o Chang (1994) y, parcialmente, con lo señalado por Weng (2004). Por el contrario, los resultados aquí analizados parecen contradecir lo indicado por Oaster (1989) o Preston

trario, bajo los supuestos del modelo AFI, sólo se observan diferencias estadísticamente significativas para la relación entre NPO-2op y NPO-5op con N del NEO-FFI o para la relación entre NPO-2op y NPO-5op con C del NEO-FFI. Finalmente, bajo los supuestos del modelo TRI no se observan diferencias estadísticamente significativas para ninguna de las relaciones analizadas.

y Colman (2000), ya que éstos investigadores sugieren que la estabilidad temporal de las escalas de personalidad aumenta a medida que incrementa el número de opciones de respuesta. En todo caso, conviene tener presente que los estudios sobre la fiabilidad test-retest realizados por éstos últimos autores se realizan sobre la base de una temporalidad mayor, entre la primera y segunda aplicación del mismo test, a la utilizada en este trabajo. De esta forma, Oaster (1989) deja pasar entre 2 a 7 días para aplicar por segunda vez el mismo cuestionario, mientras que Preston y Colman (2000) utilizan un lapso que varía entre 1 y 3 semanas. Así, el hecho de que en este trabajo se haya utilizado un test-retest *intra-sesión*, con un lapso aproximado de 45 minutos a 1 hora entre la primera y segunda aplicación de la escala SWLS, puede estar influyendo sobre los resultados reportados.

Adicionalmente, el examen de la fiabilidad test-retest de la escala SWLS sugiere que los sujetos muestran cierto grado de inconsistencia al responder la misma escala, en al menos dos ocasiones diferentes. Aunque elevada, la correlación test-retest de ninguna de las condiciones experimentales analizadas es significativamente igual a uno.

En su conjunto, los resultados reportados en este trabajo respecto del efecto del número de opciones de repuesta sobre la estabilidad temporal de la escala SWLS deben ser tomados con precaución, en tanto se basan en un análisis de fiabilidad test-retest *intra-sesión* y, al mismo tiempo, no se han podido contrastar con otra escala de personalidad (p. ej., la escala NPO). Así, se sugiere llevar a cabo nuevas investigaciones sobre la materia, incluyendo más de una escala de personalidad y un lapso test-retest mayor (p. ej., una semana o un mes).

7. Conclusiones

De acuerdo a los resultados analizados en el capítulo 5 y la discusión sobre los mismos hecha en el capítulo 6, es posible señalar que el número de opciones de respuesta parece afectar de forma leve a moderada aspectos tales como la *bondad de ajuste* (estructura interna) y, en función del modelo psicométrico utilizado como referente (TCT, AFI o TRI), la *consistencia interna* de las escalas de personalidad evaluadas en este trabajo de Tesis de Doctorado. Por el contrario, el número de alternativas de respuesta parece no afectar la *correlación entre las distintas condiciones experimentales propuestas y sus respectivas variables criterio* (evidencia basada en la relación con otras variables) o la *estabilidad temporal* (fiabilidad test-retest) de las escalas analizadas.

Tomados en su conjunto, los resultados analizados en este trabajo de Tesis de Doctorado sugieren la elección de un número de opciones de respuesta en particular debe estar basado en los objetivos de la investigación que se desee realizar. Es decir, el investigador debe definir previamente si lo que desea examinar es la *precisión* (fiabilidad), la *bondad de ajuste* (evidencia basada en la estructura interna) o la *predicción de variables criterio* (evidencia basad en la relación con otras variables) de la escala de personalidad analizada. Así y de acuerdo a los distintos modelos psicométricos utilizados como referentes (TCT, AFI y TRI), es posible sugerir las siguientes conclusiones:

1. Desde el **modelo TCT**, el incremento del número de opciones de respuesta permite maximizar la *fiabilidad* del instrumento. Por el contrario, el número de opciones de respuesta parece no incidir sobre la *predicción de variables criterio* (evidencia basada en la relación con otras variables). Por lo tanto, bajo los supuestos del modelo TCT, se sugiere utilizar un

número elevado de alternativas de respuesta (p. ej., cinco opciones), con objeto de aumentar la fiabilidad del instrumento de medición de la personalidad.

2. En cuanto al **modelo AFI**, es necesario definir previamente si lo que se quiere priorizar es la *fiabilidad* o, por el contrario, la *bondad de ajuste* (evidencia basada en la estructura interna) del instrumento de medición. Así, a medida que incrementa el número de opciones de respuesta (p. ej., de dos a cinco alternativas), aumenta la fiabilidad del instrumento, en tanto disminuye la bondad de ajuste (evidencia basada en la estructura interna) del modelo que sustenta la escala. Por el contrario, si se utilizan pocas alternativas de respuesta (p. ej., dos opciones), la bondad de ajuste del modelo que fundamenta la escala gana fuerza, mientras que la fiabilidad del mismo disminuye. Por su parte, la *predicción de variables criterio* (evidencia basada en la relación con otras variables) parece no estar afectada por el número de opciones de respuesta utilizado. En definitiva, bajo los supuestos del modelo AFI, se sugiere utilizar un número elevado de alternativas de respuesta (p. ej., cinco opciones de respuesta), si lo que se desea es priorizar la fiabilidad del instrumento. Por el contrario, se recomienda utilizar pocas categorías de respuesta (p. ej., dos opciones de respuesta), si lo que se busca es maximizar la bondad de ajuste del modelo que sustenta la escala de personalidad.

3. Finalmente, desde la perspectiva del **modelo TRI**, es posible decantarse por el uso de pocas alternativas de respuesta (p. ej., dos o tres alternativas). A diferencia de lo que se observa bajo los supuestos de los modelos TCT y AFI, desde el modelo TRI no se advierte efectos del número de opciones de respuesta sobre la *precisión* (consistencia interna) de los instrumentos de medición analizados. Por el contrario, al igual que bajo los

supuestos de los modelos TCT y AFI, desde la perspectiva del modelo TRI utilizado (modelo de respuesta graduada de Samejima), a medida que incrementa el número de alternativas de respuesta (p. ej., de dos hasta cinco opciones), la *bondad de ajuste* (evidencia basada en la estructura interna) de las escalas analizadas pierde fuerza. Por su parte, la *predicción de variables criterio* (evidencia basada en la relación con otras variables) no se ve afectada por el número de opciones de respuesta utilizado. Así, bajo los supuestos del modelo TRI (modelo de respuesta graduada de Samejima), se sugiere decantarse por el uso de sólo dos o tres alternativas de respuesta, de manera tal que no se recienta innecesariamente la bondad de ajuste (evidencia basada en la estructura interna) del modelo que sustenta la escala de personalidad utilizada.

Aunque no es el objetivo de este trabajo de Tesis de Doctorado juzgar las bondades o defectos de cada uno de los modelos psicométricos utilizados como referentes (TCT, AFI o TRI), los resultados analizados parecen sugerir una clara ventaja del modelo de respuesta graduada de Samejima (modelo TRI), respecto de los modelos TCT y AFI. Como se indicó más arriba, éste modelo permite resolver la paradoja entre fiabilidad (*consistencia interna* o *precisión*) y bondad de ajuste (*evidencia basada en la estructura interna*) observada. Así, basta con utilizar pocas opciones de respuesta (p. ej., dos o tres alternativas de respuesta) para conseguir una bondad de ajuste adecuada y, al mismo tiempo, estimaciones de fiabilidad (precisión) iguales o superiores a las alcanzadas desde los modelos TCT y AFI.

Por otro lado, sin importar el modelo psicométrico utilizado, los resultados analizados en este trabajo indican que el número de opciones de respuesta afecta sustantivamente la bondad de ajuste (*evidencia basada en la estructura in-*

terna) de los modelos que sustentan las escalas NPO y SWLS. Como se señaló anteriormente, a medida que incrementa el número de opciones de respuesta, la bondad de ajuste de los modelos que sustentan las escalas antes mencionadas pierde sistemáticamente fuerza. Por ejemplo, tomando en consideración tan solo la estimación puntual del índice de bondad de ajuste aproximada RMSEA, se aprecia un ajuste adecuado para el modelo SWLS-AFI 2 opciones de respuesta (RMSEA= .06), en tanto que uno tolerable para el modelo SWLS-AFI 5 opciones de respuesta (RMSEA= .08). El caso más evidente lo representa el análisis de la escala NPO bajo los supuestos del modelo TRI, donde se aprecia un ajuste inaceptable para el modelo NPO-AFI 2 opciones de respuesta (RMSEA= .11), en tanto que uno claramente intolerable para el modelo NPO-AFI 5 opciones de respuesta (RMSEA= .26). Estos resultados sugieren realizar nuevas investigaciones que permitan clarificar esta cuestión. Una explicación preliminar de los mismos puede fundamentarse en lo señalado por Maydeu-Olivares, Kramp y García-Forero (2006). De acuerdo a estos autores, frente a un modelo mal especificado, a mayor número de opciones de respuesta es más probable detectar que el mismo es incorrecto (Maydeu-Olivares *y col.*, 2006). Es más, frente a modelos correctamente especificados, no se aprecia efectos del número de opciones de respuesta sobre la bondad de ajuste (evidencia basada en la estructura interna) de los modelos que sustentan las escalas analizadas (Maydeu-Olivares *y col.*, 2006). Así, la pérdida sistemática de fuerza para la bondad de ajuste absoluta y aproximada sobre la evidencia basada en la estructura interna de las escalas NPO y SWLS analizadas en este trabajo puede deberse al hecho antes consignado y, por tanto, no ser consecuencia directa del efecto del número de opciones de respuesta sobre dicha variable. Adicionalmente, se sugiere realizar nuevas investigaciones respecto de la validez de las escalas NPO y SWLS (principalmente, sobre la evidencia basada en la estructura interna), en tanto parece ser que no se basan en un modelo unifactorial, según lo indicado por sus respectivos autores. Sobre este mismo punto, McDonald (1999) sugiere por ejemplo que

la escala SWLS responde a un modelo bifactorial, respecto de la satisfacción con la vida *presente* y *pasada*.

Otro aspecto que requiere una especial atención es el hecho que no solo el número de opciones de respuesta parece afectar las propiedades psicométricas de los instrumentos analizados, sino también el *tiempo*. Curiosamente, al proponer un test-retest intra-sesión se esperaba observar correlaciones cercanas a uno. Esto no fue así, lo cual indica falta de consistencia en las respuestas brindadas por los individuos. Dicha falta de consistencia puede haber afectado los resultados presentados a lo largo del presente trabajo. Por tal motivo, se sugiere realizar investigaciones futuras que permitan separar el efecto *tiempo* y *número de opciones de respuesta* sobre las propiedades psicométricas de los cuestionarios de personalidad.

Como es natural, las conclusiones que se desprenden de los estudios presentados en este trabajo de Tesis de Doctorado no pueden ser generalizadas más allá de los instrumentos utilizados o de la población en la que fueron recogidos los datos. De todos modos, pensamos que el método utilizado permite generalizar en buena medida las mismas al ámbito general del diseño y aplicación de los cuestionarios para la evaluación de la personalidad. Sin perjuicio de lo anterior, se sugiere llevar a cabo nuevas investigaciones que permitan resolver cada uno de los problemas detectados y enunciados a lo largo de los capítulos 6 y 7 del presente trabajo, con especial atención en lo que se refiere al análisis del efecto del número de opciones de respuesta sobre la bondad de ajuste (*evidencia basada en la estructura interna*) de los cuestionarios diseñados para la evaluación de la personalidad.

Antes de acabar, solo resta señalar que esperamos que las conclusiones que se derivan de los estudios realizados en el presente trabajo de Tesis de Doc-

torado permitan no solo estimular nuevas investigaciones sobre el tema aquí tratado, sino también contribuyan a orientar un diseño de cuestionarios para la evaluación de la personalidad cada vez más eficaz y eficiente. Sobre esto último, naturalmente, este trabajo aborda una cuestión muy puntual sobre el diseño de los cuestionarios de personalidad. De todos modos, pensamos que es necesario aclarar definitivamente esta cuestión, en tanto parece evidente el efecto del número de opciones de respuesta sobre aspectos tales como la *bondad de ajuste* (evidencia basada en la estructura interna) y, parcialmente, la *consistencia interna* de las escalas diseñadas para la evaluación de la personalidad. En definitiva, lo anterior contribuye no solo a mejorar el diseño de cuestionarios para la evaluación de la personalidad, sino también permite contar instrumentos cada vez más fiables y válidos, orientados a la evaluación de: (a) procesos de validación o refutación de nuevas teorías, modelos o constructos psicológico que se quiera adoptar y/o (b) aspectos psicológicos que contemplen la personalidad de un individuo o grupo de individuos. Esperamos que el presente trabajo de Tesis de Doctorado contribuya, aunque solo sea de forma puntual, en un mejor desarrollo de estos tópicos.

Referencias

Abad, F. J., Olea, J. y Ponsoda, V. (2001). Analysis of the Optimum Number Alternatives from the Item Response Theory. *Psicothema, 13*(1), 152-158.

Aiken, L. R. (1983). Number of response categories and statistics on a teacher rating scale. *Educational and Psychological Measurement, 43*, 397-401.

Allen, M. J. y Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

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

Andrés, A. (1997). *Manual de Psicología Diferencial*. Madrid: McGraw-Hill / Interamericana de España.

American Psychological Association, American Educational Research Association y National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washinton, DC: American Psychological Association.

Arrinddell, W., Meeuwesen, L. y Huyse, F. (1991). The Satisfaction With Life Scale: Psychometric properties in a non-psychiatric medical outpatients sample. *Personality and Individual Differences, 12*(2), 117-123.

Bandalos, D. L. y Enders, C. K. (1996). The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education, 9*(2), 151-160.

Bartholomew, D. J. (1995). Spearman and the origin and development of factor analysis. *British Journal of Mathematical and Statistical Psychology, 48*, 211-220.

Bearden, W. O., Netmeyer, R. G. y Mobley, M. F. (1993). *Handbook of marketing scales: multi-item measures for marketing and consumer behaviour research.* Newbury Park, CA: Sage.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modelling. *Annual Review of Psychology, 31*, 238-246.

Bentler, P. M. (1986). Structural modelling and Psychometrika: An historical perspective on growth and achievements. *Psychometrika, 51*, 35-51.

Bentler, P. M. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M. (1995). EQS Structural Equations Program Manual (Version 6.1). Encino, CA: Multivariate Software, Inc.

Berry, J. W. (1980). Introduction to methodology. En H. Triandis y J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (Vol. 2, pp. 1-28). Boston: Allyn & Bacon.

Birnbaum, A. (1957). *Efficient design and use of tests of a mental ability for various decision-making problems*. Texas: USAF School of Aviation Medicine, Randolph Air Force Base.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F. M. Lord y M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A. y Long, J. S. (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Brislin, R. W. (1980). Translation and content analysis of oral and written materials. En H. Triandis y J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (Vol. 2, pp. 389-444). Boston: Allyn & Bacon.

Browne, M. W. (1992). FITMOD: Point and interval estimates of measures of fit of a model. Ohio: The Ohio State University.

Browne, M. W. y Cudeck, R. (1993). Alternative ways of assessing model fit. En K. A. Bollen y J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Bruno, J. E. y Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information Theoretic Perspective. *Educational and Psychological Measurement, 55*(6), 959-966.

Calero, M. D. y Padilla, J. L. (2004). Técnicas psicométricas: los tests. En R. Fernández-Ballesteros (Ed.), *Evaluación psicológica. Conceptos, métodos y estudios de casos.* Madrid: Pirámide.

Campbell, N. R. (1938). *Symposium: Measurement and its importance for philosophy* (Vol. 17, Suplemento). Londres: Harrison.

Cicchetti, D. V., Showalter, D. y Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of inter-rater reliability: A Monte-Carlo investigation. *Applied Psychological Measurement, 9*, 31-36.

Comrey, A. L. y Montag, I. (1982). Comparison of factor analytic results with two choice and seven choice personality item formats. *Applied Psychological Measurement, 6*(3), 285-289.

Coombs, C. H. (1964). *A theory of data*. New York: John Wiley.

Costa, P. T. y McCrae, R. R. (1992). *Professional Manual. Revised NEO Personality Inventory (NEO-PI-R) and NEO-Five Factor Inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T. y McCrae, R. R. (1999). *Manual profesional del Inventario de Personalidad NEO Revisado (NEO-PI-R) y del Inventario NEO reducido de Cinco Factores (NEO-FFI)*. Madrid: Tea Ediciones.

Costin, F. (1970). The optimal number of alternatives in multiple-choice achievement tests: Some empirical evidence for a mathematical proof. *Educational and Psychological Measurement, 30*, 353-358.

Costin, F. (1972). Three-choice versus four-choice items: implications for reliability and validity of objective achievement test. *Educational and Psychological Measurement, 32*, 1035-1038.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*(4), 407-422.

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

Cronbach, L. J. (1990). *Essentials of psychological testing* (5 ed.). New York: Harper & Row.

Champney, H. y Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology, 23*, 323-331.

Chang, L. (1994). A psychometric evaluation of four-point and six-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*, 205-215.

Churchill, G. A., Jr. y Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research, 21*(4), 360-375.

D'Zurilla, T. J. y Chang, E. C. (1995). The relations between social problem solving and coping. *Cognitive Therapy and Research, 19*, 547-562.

D'Zurilla, T. J. y Maydeu-Olivares, A. (1995). Conceptual and methodological issues in social problem-solving assessment. *Behavior Therapy, 26*, 415-438.

D'Zurilla, T. J. y Nezu, A. M. (1999). *Problem-Solving Therapy* (Second ed.). New York: Springer.

D'Zurilla, T. J., Nezu, A. M. y Maydeu-Olivares, A. (2002). *The Social Problem Solving Inventory - Revised (SPSI-R): Technical manual*. North Tonawanda, NY: Multi-Health Systems, Inc.

Diener, E., Emmons, R. A., Larsen, R. J. y Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71-75.

Eysenck, H. J. y Eysenck, S. B. G. (1969). *Personality structure and measurement*. San Diego: Knapp.

Eysenck, H. J. y Eysenck, S. B. G. (1978). *EPQ cuestionario de personalidad*. Madrid: TEA.

Fernández-Ballesteros, R. (2004). *Evaluación psicológica. Conceptos, métodos y estudio de casos*. Madrid: Pirámide.

Ferrando, P. J. (1995). Equivalencia entre los formatos Likert y continuo en ítems de personalidad: Un estudio empírico. *Psicológica, 16*, 417-428.

Ferrando, P. J. (1996). Relaciones entre el análisis factorial y la teoría de respuesta a los ítems. En J. Muñiz (Ed.), *Psicometría* (pp. 555-612). Madrid: Editorial Universitaria.

Ferrando, P. J. (1999). Likert scaling using continuos, censored, and graded response models: Effects on criterion-related validity. *Applied Psychological Measurement, 23*(2), 161-175.

Ferrando, P. J. (2000). Testing the equivalence among different item response formats in personality measurement: A structural equation modelling approach. *Structural Equation Modeling, 7*(2), 271-286.

García-Cueto, E., Muñiz, J. y Lozano, L. M. (2002). Influencia del número de alternativas en las propiedades psicométricas de los tests. *Metodología de las Ciencias del Comportamiento, volumen especial*.

García-Cueto, E., Muñiz, J. y Lozano, L. M. (2003). *Efecto de la reducción de alternativas en las escalas tipo likert desde la perspectiva de la TRI (Effect of the reduction of alternatives in a Likert-type scale under the IRT perspective).* Paper presentado en la Conferencia IX Conferencia Española de Biometría, La Coruña (Spain), may 28-30th.

Garner, W. R. (1960). Rating scales, discriminability and information transmission. *Psychological Review, 67*(6), 343 - 352.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica, 40*, 979-1001.

Gómez, J. (1996). Aportaciones de los modelos de estructuras de covariancia al análisis psicométrico. En J. Muñiz (Ed.), *Psicometría* (pp. 457-554). Madrid: Editorial Universitaria.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement., 12*(2), 109-113.

Grier, J. B. (1976). The optimal number of alternatives at a choice point with travel time considered. *Journal of Mathematical Psychology, 12*, 31-97.

Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum.

Hair, J. F., Anderson, R. E., Tatham, R. L. y Black, W. (1999). *Análisis multivariante* (5ª ed.). Madrid: Prentice Hall Ibérica.

Hambleton, R. K. y Swaminathan, H. (1985). *Item Response Theory. Principles and Applications*. Boston: Kluwer-Nijhoff.

Hernández, A., Muñiz, J. y García-Cueto, E. (2000). Comportamiento del modelo de respuesta graduada en función del número de categorías de la escala. *Psicothema, 12*(Supl. nº 2), 288-291.

HU, L. T. y Bentler, P. M. (1995). Evaluating model fit. En R. H. Hoyle (Ed.), *Structural equation modeling* (pp. 76-99). Thousand Oaks, CA: Sage.

Jenkins, G. D., Jr. y Taber, T. D. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62*, 392-398.

John, O. P. (1990). The "big five" factor taxonomy: Dimensions of personality in the natural language and questionnaires. En L. A. Pervin (Ed.), *Handbook of Personality. Theory and Research* (pp. 66-100). New York: Guilford Press.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. En A. S. Goldberger y O. D. Duncan (Eds.), *Structural Equation Models in the Social Sciences* (pp. 85-112). New York: Academic Press.

Jöreskog, K. G. y Söborn, D. (1979). *Advances in factorial analysis and structural equation models*. Cambridge, MA: Abt Books.

Jöreskog, K. G. y Sörbom, D. (2005). Lisrel (Version 8.57). Chicago: Scientific Software International, Inc.

Keesling, J. W. (1972). *Maximum Likelihood Approaches to Causal Analysis.* Phd dissertation. University of Chicago, Chicago.

Komorita, S. S. y Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement, 25*(4), 987-995.

Kramp, U., Maydeu-Olivares, A., Tous, J. M. y Gallardo, D. (2004). Coping Strategy Indicator: Reliability and validity of the Spanish adaptation and its relationship with personality and affectivity dispositions. *Poster presented in the 7th Working Day of the Society for the Individual Differences Assessment (Sociedad para la Investigación de las Diferencias Individuales: SEIDI, 17 th September*), Lleida (Spain).

Krane, W. R. y Slaney, K. L. (2005). A general introduction to the common factor model. En Maydeu-Olivares y J. J. McArdle (Eds.), *Contemporary Psychometrics. A Festschrift for Roderick P. McDonald* (pp. 125-151). Mahwah, NJ: Lawrence Erlbaum Associates.

Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh A, 60*, 64-82.

Lawley, D. N. (1943a). The application of maximum likelihood method to factor analysis. *British Journal of Psychology, 33*, 172-175.

Lawley, D. N. (1943b). On problems connected with item selection and test construction (Vol. 61-A, pp. 273-287): Proceedings of the Royal Society of Edinburgh.

Lawley, D. N. (1944). The Factorial Analysis of multiple item test (Vol. 62-A, pp. 74-82): Proceedings of the Royal Society of Edinburgh.

Lawley, D. N. y Maxwell, A. E. (1971). *Factor analysis as a statistical method*. New York: Elsevier.

Lazarsfeld, P. F. (1959). Latent structure analysis. En S. Koch (Ed.), *Psychology: A study of a science* (Vol. 3). New York: McGraw Hill.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 5-55.

Lissitz, R. W. y Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60*, 10-13.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs, núm. 7*.

Lord, F. M. (1953a). An application of confidence intervals of maximum likehood to the estimation of an examinee's ability. *Psychometrika, 18*, 57-75.

Lord, F. M. (1953b). On the statistical treatment of football numbers. *The American Psychologist, 8*, 750-751.

Lord, F. M. (1953c). The relation of test scores to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-549.

Lord, F. M. (1977). Optimal number of choices per item: A comparison of four approaches. *Journal of Educational Measurement., 14*(1), 33-38.

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: LEA.

Lord, F. M. y Novick, M. R. (1968). *Statistical theories of mental tests scores*. Reading, MA.: Addison-Wesley.

Luce, R. D. y Tukey, J. W. (1964). Simultaneous conjoint measurement: A new form of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1-27.

Martínez-Arias, R. (1996). *Psicometría: Teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Masters, J. R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement, 11*(1), 49-53.

Matell, M. S. y Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement, 31*, 657-674.

Matell, M. S. y Jacoby, J. (1972). Is there an optimal number of alternatives for Likert scale items? *Journal of Applied Psychology, 56*, 506-509.

Maydeu-Olivares, A. (1991). *Un modelo de ecuaciones estructurales de la resolución de problemas sociales: Efectos de la aplicación de la Teoría de Respuesta a los Ítems.* Universitat de Barcelona, Barcelona.

Maydeu-Olivares, A. (2005a). Further Empirical Results on Parametric Versus Non-Parametric IRT Modeling of Likert-Type Personality Data. *Multivariate Behavioral Research, 40(2)*, 261-279.

Maydeu-Olivares, A. (2005b). Linear Item Response theory, Nonlinear Item Response theory, and Factor Analysis: A Unified Framework. En Maydeu-Olivares y J. J. McArdle (Eds.). *Contemporary Psychometrics. A Festschrift for Roderick P. McDonald* (pp. 73-100). Mahwah, NJ: Lawrence Erlbaum Associates.

Maydeu-Olivares, Kramp, U. y García-Forero, C. (2006, en preparación). Effects of number of response categories on power to detect model mispecification in item factor analysis. *Psychological Methods*.

Maydeu-Olivares, A. y Coffman, D. L. (2005). *An asymptotically distribution free (ADF) interval estimator for coefficient alpha.* Paper presentado en la Conferencia Annual Meeting of the Society of Multivariate Experimental Psychology, Tahoe City (USA).

Maydeu-Olivares, A. y D'Zurilla, T. J. (1995). A factor analysis of the Social Problem-Solving Inventory using polychoric correlations. *European Journal of Psychological Assessment, 11*, 98-107.

Maydeu-Olivares, A. y D'Zurilla, T. J. (1996). A factor analytic study of the Social Problem-Solving Inventory: An integration of theory and data. *Cognitive Therapy and Research, 20*, 115-133.

Maydeu-Olivares, A., Gallardo, D. y Kramp, U. (2004). *Parametric vs. non-parametric IRT modeling of Likert-type personality data.* Paper presentado en la Conferencia VII Jornada de la Sociedad Española para la Investigación en Diferencias Individuales (SEIDI), Lleida (Spain).

Maydeu-Olivares, A. y McArdle, J. J. (2005). *Contemporary Psychometrics. A Festschrift for Roderick P. McDonald*. Mahwah, NJ: Lawrence Erlbaum Associates.

Maydeu-Olivares, A., Rodríguez-Fornells, A., Gómez-Benito, J. y D'Zurilla, T. J. (2000). Psychometric properties of the Spanish adaptation of the Social Problem-Solving Inventory-Revised (SPSI-R). *Personality and Individual Differences, 29*, 699-708.

McCallum, D. M., Keith, B. R. y Wiebe, D. J. (1988). Comparison of response formats for multidimensional health locus of control scales: Six levels versus two levels. *Journal of Personality Assessment, 52*(4), 732-736.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillside, NJ: Lawrence Erlbaum Associates.

McDonald, R. P. (1997). Normal ogive multidimensional model. En W. J. Van der Linden y R. K. Hambleton (Eds.), *Handbook of item response theory* (pp. 258-269). New York: Springer-Verlag.

McDonald, R. P. (1999). *Test Theory. A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-hill.

Muñiz, J. (1996). *Psicometría*. Madrid: Editorial Universitaria.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Madrid: Pirámide.

Muñiz, J. (2003). *Teoría Clásica de los Test*. Madrid: Pirámide.

Muñiz, J., García-Cueto, E. y Lozano, L. M. (2005). Item format and the psychometric properties of the Eysenck Personality Questionnaire. *Personality and Individual Differences, 38*, 61-69.

Muraki, E. (1990). Fitting a Polytomous Item Response Model to Likert-Type Data. *Applied Psychological Measurement, 14*(1), 59-71.

Murphy, K. R. y Davidshofer, C. O. (1994). *Psychological testing: Principles and applications* (Third ed.). Englewood Cliffs, NJ: Prentice-Hall.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49*, 115-132.

Muthén, B. O. (1993). Goodness of fit with categorical and other nonnormal variables. En K. A. Bollen y J. S. Long (Eds.), *Testing structural equation models* (pp. 205-234). Newbury Park CA: Sage.

Muthén, B. O., du Toit, S. H. C. y Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes: *Artículo no publicado*.

Muthén, L. K. y Muthén, B. O. (2001). *Mplus user's guide* (Second ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, L. K. y Muthén, B. O. (2004). *Mplus user's guide. Statistical analysis with latent variables* (Second ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. y Muthén, L. K. (2005a). Chi-Square Difference Testing Using the Satorra-Bentler Scaled Chi-Square. Visto el 06 de noviembre, desde http://www.statmodel.com/chidiff.shtml.

Muthén, L. K. y Muthén, B. O. (2005b). Mplus (Versión 3.13). Los Angeles, CA: Muthén & Muthén.

Neto, F. (1993). The Satisfaction With Life Scale: Psychometric properties in an adolescent sample. *Journal of Youth Adolescence, 22*(2), 125-134.

Nezu, A. M. y D'Zurilla, T. J. (1989). Social problem solving and negative affective conditions. En P. C. Kendall y D. Watson (Eds.), *Anxiety and depression: Distinctive and overlapping features* (pp. 285-315). New York: Academic Press.

Nunnally, J. C. (1967). *Psychometric Theory*. New York: McGraw-Hill.

Oaster, T. R. F. (1984a). Coefficient alpha reliability of Likert-type scales as a function of the number of alternatives per choice point: Manuscrito no publicado, University of Missouri, Kansas City, MO.

Oaster, T. R. F. (1984b). *Even numbered alternative scale reliability.* Paper presentado en la Conferencia Annual meeting of the American Educational Research Association, New Orleans, Louisiana.

Oaster, T. R. F. (1985). *Stability of Likert-type scales with varying numbers of alternatives per choice point.* Paper presentado en la Conferencia Annual meeting of the American Educational Research Association, Chicago, Illinois.

Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-type scales. *Perceptual and Motor Skills, 68*, 549-550.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443-460.

Pavot, W. y Diener, E. (1993). Review of the Satisfaction With Life Scale. *Psychological Assessment, 5*(2), 164-172.

Pavot, W., Diener, E., Colvin, C. R. y Sandvik, E. (1991). Further validation of the Satisfaction with Life Scale: evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment, 57*(1), 149-161.

Peabody, D. (1962). Two components in bipolar scales: Direction and extremeness. *Psychological Review, 69*, 65-73.

Pelechano, V. (2000). *Psicología sistemática de la personalidad*. Barcelona: Ariel.

Pemberton, H. E. (1933). A technique for determining the optimal rating scale for opinion measures. *Sociology and Social Research, 17*, 470-472.

Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research, 16*(February), 6 - 17.

Pivot, W., Diener, E., Colvin, C. R. y Sandvik, E. (1991). Further validation of the satisfaction with life scale: Evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment, 57*(1), 149-161.

Preston, C. C. y Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* (104), 1-15.

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika, 38*, 513-533.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedagogiske Institut.

Renom, J. (1992). *Diseño de Tests*. Barcelona: IDEA I+D.

Richardson, M. W. (1936). Notes of the rationale of item analysis. *Psychometrika, 1*(1), 69-76.

Rogers, W. T. y Harley, D. (1999). An empirical comparison of three- and four-choice items and test: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement, 59*(2), 234-247.

Royce, J. y Powell, A. (1983). *Theory of personality and individual differences. Factors, systems and processes*. Englewood Cliffs, NJ: Prentice-Hall.

Ruch, G. M. y Charles, J. W. (1928). A comparison of five types of objective tests in elementary psychology. *Journal of Applied Psychology, 12*, 398-404.

Ruch, G. M., DeGraff, M. H. y Gordon, W. E. (1926). *Objective examination methods in the social studies*. New York: Scott, Foresman and Co.

Ruch, G. M. y Stoddard, G. D. (1925). Comparative reliabilities of five types of objective examinations. *Journal of Educational Psychology, 16*, 89-103.

Ruch, G. M. y Stoddard, G. D. (1927). *Tests and measurements in high school instruction*. Chicago: World Book.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement, 17*.

Sancerni, M. D., Meliá, J. L. y González, V. (1990). Formato de respuesta, fiabilidad y validez, en la medición del conflicto de rol. *Psicológica, 11*, 167-175.

Sandin, B., Chorot, P., Lostao, L., Joiner, T. E., Santed, M. A. y Valiente, R. M. (1999). Escalas Panas de afecto positivo y negativo: Validación factorial y convergencia transcultural. *Psicothema, 11*(1), 37-51.

Santisteban-Requena, C. (1990). *Psicometría. Teoría y práctica en la construcción de tests*. Madrid: Ediciones Norma.

Satorra, A. y Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. En A. von Eye y C. C. Clogg (Eds.),

Latent variable analysis: Applications to developmental research (pp. 399-419). Thousand Oaks, CA: Sage.

Satorra, A. y Bentler, P. M. (2001). A scaled difference Chi-Square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507-514.

Shaw, M. E. y Wright, J. M. (1967). *Scales for the measurement of attitudes.* New York: McGraw-Hill.

Shevlin, M., Brunsden, V. y Miles, J. (1998). Satisfaction With Life Scale: Analysis of factorial invariance, mean structures and reliability. *Personality and Individual Differences, 25*(5), 911-916.

Spearman, C. (1904a). 'General Intelligence' objectively determined and measured. *American Journal of Psychology, 15*, 202-293.

Spearman, C. (1904b). The proff and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology, 3*, 271-295.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5*, 417-426.

Spearman, C. (1927). *The Abilities of Man*. Londres: McMillan.

SPSS. (2005). Amos (Version 6.0). Chicago: SPSS Inc.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173-180.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 667-680.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. En S. S. Stevens (Ed.), *Handbook of Experimental Psychology* (pp. 1-49). New York: Willey.

Straton, R. G. y Catts, R. M. (1980). A comparison of two, three and four-choice item test given a fixed total number of choices. *Educational and Psychological Measurement, 40*, 357-365.

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology, 7*(December), 456-461.

Thurstone, L. L. (1931). *The reliability and validity of tests*. Ann Arbor, MI: Edwards brothers.

Thurstone, L. L. (1947). *Multiple factor analysis: a development and extension*. Chicago: University Press.

Toops, H. A. (1921). Trade tests in education. *Teachers College Contributions to Education, 115*.

Torgerson, W. S. (1962). *Theory and methods of scaling* (Third ed.). London: John Wiley & Sons.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1-13.

Tucker, L. R. y Lewis, C. (1973). The Reliability Coefficient for Maximum Likelihood Factor Analysis. *Psychometrika, 38*, 1-10.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology, 1*, 386-391.

Van der Linden, W. J. y Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Velicer, W. F. y Stevenson, J. F. (1978). The relation between item format and the structure of the Eysenck Personality Inventory. *Applied Psychological Measurement, 2*(2), 293-304.

Wainer, H. y Barun, H. I. (1988). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Walsh, W. B. y Betz, N. E. (1995). *Tests and Assessment* (Third ed.). Englewood Cliffs, NJ: Prentice-Hall.

Watson, D., Clark, A. y Tellegen, A. (1988). Development and validation of brief measure of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956-972.

Widiger, T. A. y Trull, T. J. (1997). Assessment of the five-factor model of personality. *Journal of Personality Assessment, 68*(2), 228-250.

Wiley, D. E. (1973). The identification problem for structural equation models with unmeasured variables. En A. S. Goldberger y O. D. Duncan (Eds.), *Structural Equation Models in the Social Sciences* (pp. 69-83). New York: Academic Press.

Wolfram Research, I. (2003). Mathematica (Version 5.0.1.0). Champaign, IL: Wolfram Research, Inc.

Yuan, K., Guarnaccia, C. y Hayslip, B. (2003). A study of the distribution of sample coefficient alpha with the Hopkins symptom checklist: Bootstrap versus asymptotics. *Educational and Psychological Measurement, 63*(1), 5-23.

Anexos

Índice de anexos

Anexo 1: Adaptación al castellano de la escala PANAS utilizada.

A continuación se reproduce la adaptación al castellano del cuestionario *Positive Affect and Negative Afect Schedule* (PANAS) utilizado en este trabajo de investigación. Cabe señalar que la escala de respuesta fue modificada siguiendo los criterios expuestos en el subapartado (4.3).

Instrucciones

Esta escala consiste de una serie de palabras que describen diferentes sentimientos y emociones. Lea cada ítem e indique cómo se siente habitualmente utilizando la siguiente escala.

***Insertar escala de respuesta que corresponda* (2, 3 ó 5 opciones de respuesta)**

1. Interesado/a
2. Angustiado/a
3. Emocionado/a
4. Molesto/a
5. Fuerte
6. Culpable
7. Asustado/a
8. Hostil
9. Entusiasta
10. Orgulloso/a
11. Irritable
12. Alerta
13. Avergonzado/a
14. Inspirado/a
15. Nervioso/a
16. Decidido/a
17. Atento/a
18. Inquieto/a
19. Activo/a
20. Temeroso/a

Anexo 2: Adaptación al castellano de la escala SWLS utilizada.

A continuación se reproduce la adaptación al castellano del cuestionario *Satisfation with Life* (SWLS) utilizado en este trabajo de investigación. Cabe señalar que la escala de respuesta fue modificada siguiendo los criterios expuestos en el subapartado (4.3). Asimismo, los reactivos destacados en negrita corresponden a los ítems utilizados en los distintos análisis realizados, en tanto que los reactivos restantes sirven como distractores.

Instrucciones

Por favor, señale en qué medida está usted de acuerdo o en desacuerdo con cada una de las preguntas que se formulan más abajo. Lea cuidadosamente cada una de ellas antes de contestar y responda, según sea el caso, utilizando la escala que aparece a continuación.

***Insertar escala de respuesta que corresponda* (2, 3 ó 5 opciones de respuesta)**

1. **En muchos sentidos mi vida está cerca de mi ideal.**
2. Frecuentemente pienso sobre tiempos infelices o hechos de mi pasado.
3. **Las condiciones que rodean mi vida son excelentes.**
4. Me agrada la vida que he tenido.
5. **Estoy satisfecho con mi vida.**
6. Frecuentemente experimento emociones negativas (desagradables) que me hacen infeliz.
7. **Hasta el momento he conseguido las cosas importantes que quería en la vida.**
8. Cuando algo me hace feliz, dicha emoción dura generalmente mucho tiempo.
9. **Si pudiese volver a vivir mi vida, no cambiaría nada.**
10. Mi vida no sigue los estándares de lo que yo considero debe ser una buena vida (una vida ideal).

Anexo 3: Escala NPO utilizada.

A continuación se reproduce la escala *Orientación Negativa a los Problemas* (NPO) utilizado en este trabajo de investigación. Cabe señalar que la escala de respuesta fue modificada siguiendo los criterios expuestos en el subapartado (4.3).

Instrucciones

Las siguientes frases describen la forma en que algunas personas piensan, sienten o se comportan cuando se enfrentan a problemas que aparecen en su vida. Nos referimos a aquellos problemas que pueden tener un efecto significativo en su bienestar o en el bienestar de las personas que aprecia. Por favor lea cada una de estas frases y escoja la respuesta que considere refleja de forma mejor la forma en que piensa, siente y se comporta cuando se enfrenta a problemas que han surgido más recientemente en su vida.

***Insertar escala de respuesta que corresponda* (2, 3 ó 5 opciones de respuesta)**

1. Pierdo demasiado tiempo preocupándome de mis problemas en vez de intentar resolverlos.

2. Habitualmente me siento amenazado/a y temeroso/a cuando tengo un problema importante que resolver.

3. A menudo me siento nervioso/a y poco seguro/a de mi mismo/a cuando tengo una decisión importante que tomar.

4. Cuando mis primeros intentos de resolver un problema fracasan, me enfado y me frustro mucho.

5. Cuando me enfrento a un problema difícil, a menudo dudo que sea capaz de resolverlo yo sólo, no importa lo mucho que lo intente.

6. Me molestan mucho los problemas difíciles.

7. Cuando estoy intentando resolver un problema, a menudo me altero tanto que no puedo pensar fríamente.

8. Odio tener que resolver los problemas que ocurren en mi vida.

9. Cuando tengo un problema importante que resolver, a menudo me deprimo y soy incapaz de hacer nada.

10. Cuando mis primeros intentos de resolver un problema fracasan, tiendo a desanimarme y deprimirme.

Anexo 4: Efecto del número de opciones de respuesta sobre los estadísticos descriptivos de las escalas NPO y SWLS, bajo los supuestos del modelo de Teoría Clásica de los Tests (TCT).

El anexo 4 analiza el efecto del número de opciones de respuesta sobre los estadísticos descriptivos de las escalas NPO y SWLS, bajo los supuestos del modelo de Teoría Clásica de los Tests (TCT). Específicamente, este análisis consiste en: (a) examinar la relación entre las medias ($\overline{X}$) de las distintas condiciones experimentales, por medio de una *Prueba T de Student* para muestras relacionadas, y (b) comparar las desviaciones estándar (*D.S.*) observadas para las distintas condiciones experimentales. En caso de encontrarse diferencias estadísticamente significativas para la relación entre las $\overline{X}$ estimadas para las distintas condiciones experimentales, se concluye que el número de opciones de respuesta afecta los estadísticos descriptivos de las escalas NPO y SWLS. Por el contrario, si no se observan diferencias estadísticamente significativas para la relación entre las $\overline{X}$ de las distintas condiciones experimentales, se concluye que la variación del número de alternativas de respuesta no afecta los estadísticos descriptivos de las escalas bajo estudio.

La tabla 22 muestra los estadísticos descriptivos de las escalas NPO y SWLS, desde la óptica del modelo TCT. Como puede observarse en dicha tabla, las $\overline{X}$ y las *D.S.* de ambas escalas aumentan sistemáticamente a medida que incrementa el número de opciones de respuesta. Tras compararse las $\overline{X}$ estimadas para las distintas condiciones experimentales (dos, tres y cinco opciones de respuesta) y escalas (NPO y SWLS) por medio de una *Prueba T de Student* para muestras relacionadas (*t*), se comprueba que las diferencias observadas para la relación entre las $\overline{X}$ de las distintas condiciones experimentales, tanto para la escala NPO como para la escala SWLS, son estadísticamente significativa (p<

.01). Así, es posible concluir que el número de alternativas de respuesta afecta la expresión de los estadísticos descriptivos de las escalas analizadas, en la siguiente dirección: a medida que aumenta el número de opciones de respuesta, incrementa tanto la $\overline{X}$ como la dispersión (*D.S.*) de las escalas NPO y SWLS. Estos resultados son consistentes con lo sugerido por Aiken (1983) y Muñiz, García-Cueto y Lozano (2005). De acuerdo a lo señalado por Aiken (1983), a medida que incrementa el número de opciones de respuesta, aumenta la $\overline{X}$ de las escalas de personalidad. Por su parte, Muñiz, García-Cueto y Lozano (2005) sugieren conclusiones similares, aunque respecto al comportamiento de la *D.S.* de las escalas de personalidad: a medida que aumenta el número de opciones de respuesta, incrementa la *D.S.* de las mismas.

Tabla 22: Estadísticos descriptivos y Prueba T de Student para las escalas NPO y SWLS, desde el modelo de Teoría Clásica de los tests (TCT).

Estadísticos descriptivos				
Número de opciones de respuesta	NPO		SWLS	
	$\overline{X}$	D.S.	$\overline{X}$	D.S.
2	3.86	2.75	3.11	1.44
3	7.76	4.51	6.06	2.33
5	16.37	6.52	11.44	3.36
	Prueba T			
Relación entre formatos de respuesta	*t*	*gl*	*t*	*gl*
2 y 3	- 40.73**	745	- 40.72**	425
2 y 5	- 76.65**	745	- 73.09**	425
3 y 5	- 68.57**	745	- 61.57**	425
N	746		426	

Nota: $\overline{X}$ = Media; *D.S.* = Desviación Estándar; *t* = valor de la prueba T de Student; *gl* = Grados de libertad; N = Tamaño de la muestra. NPO = *Orientación Negativa a los Problemas*; SWLS = *Escala de Satisfacción con la Vida*.
** $p < .01$.

Anexo 5: Efecto del número de opciones de respuesta sobre el porcentaje de varianza explicado por las escalas NPO y SWLS, utilizando Análisis de Componentes Principales (ACP)

El anexo 5 presenta el análisis del efecto del número de opciones de respuesta sobre el porcentaje de varianza explicado por las escalas NPO y SWLS, utilizando Análisis de Componentes Principales (ACP). Se procede a examinar el porcentaje de varianza explicado por el factor evaluado por las escalas NPO y SWLS, en función del número de opciones de respuesta utilizado. De esta forma, si el porcentaje de varianza explicado por el factor evaluado por las escalas NPO o SWLS no varía entre las distintas condiciones experimentales examinadas, se concluye que el número de opciones de respuesta no afecta ésta condición. Por el contrario, si el porcentaje de varianza explicado varía entre las distintas condiciones experimentales, se concluye que el número de categorías de respuesta si afecta la expresión del factor evaluado por las escalas experimentales (NPO o SWLS).

La tabla 23 muestra el porcentaje de varianza explicado por el factor evaluado por las escalas NPO y SWLS, en función del número de opciones de respuesta utilizado para su estimación. Los resultados observados en dicha tabla sugieren el porcentaje de varianza explicado, por cada condición experimental (dos, tres y cinco opciones de respuesta) y escala (NPO y SWLS), parece incrementar a medida que aumenta el número de opciones de respuesta. Así, para la escala NPO varía desde un 34.37 (NPO 2 op) hasta un 49.61 (NPO 5op) por ciento, en tanto que para la escala SWLS varía desde un 40.95 (SWLS 2op) hasta un 52.31 (SWLS 5op) por ciento.

Tabla 23: Porcentaje de varianza explicada por el primer factor de las escalas NPO y SWLS.

	NPO			SWLS		
	Formato de respuesta			Formato de respuesta		
	2	3	5	2	3	5
% de varianza	34.37	40.90	49.61	40.95	47.35	52.31

Nota: NPO = *Orientación Negativa a los Problemas*; SWLS = *Escala de Satisfacción con la Vida*.

En su conjunto, los resultados observados en la tabla 23 indican que el número de opciones de respuesta parece afectar el porcentaje de varianza explicado por el factor evaluado por las escalas NPO y SWLS. Así, a medida que aumenta el número de alternativas de respuesta, incrementa también el porcentaje de varianza explicado. Estos resultados son consistentes con lo señalado por Velicer y Stevensons (1978), García-Cueto, Muñiz y Lozano (2002) y Muñiz, García-Cueto y Lozano (2005).