Advances in Computerized Adaptive Measurement of Personality

by

María Dolores Nieto

A doctoral dissertation submitted
to the Faculty of Psychology
in Universidad Autónoma de Madrid
in partial fulfillment for the degree of Doctor of Philosophy

Directors:
Dr. Francisco J. Abad and Dr. Luis Eduardo Garrido

July 2019
Madrid, Spain

This dissertation is in memoriam of Julio Olea.

None of this would be possible without him.

Nothing in life is to be feared, it is only to be understood.

Now is the time to understand more, so that we may fear less.

MARIE CURIE

A la memoria de Julio Olea

Gran parte de esta tesis ha sido posible gracias a la constancia e ilusion de Julio Olea (1961-2018), catedrático del Departamento de Psicología Social y Metodología en la Facultad de Psicología de la Universidad Autónoma de Madrid.

Aunque son y serán varios los homenajes que se harán en recuerdo de Julio, quisiera que esta tesis conste como mi pequeño homenaje personal a la memoria de mi codirector, el cual por motivos de salud falleció el pasado 1 de octubre del 2018. Julio ha sido uno de los principales impulsores de muchas de las ideas que forman parte de esta tesis desde sus comienzos. En particular, su trabajo en los dos primeros estudios desarrollados, los cuales ya han sido publicados, ha sido fundamental para llevar esta tesis a buen puerto. Pero más allá de las formalidades académicas que se destacarán en otros actos, quisiera recordar mi experencia como doctoranda junto a Julio desde un punto de vista más personal.

Julio era un hombre con una personalidad muy particular. Todo aquel que le conocía era incapaz de no apreciarle. En general, si hubiera que describirle en una sola frase, podría decirse que era un hombre exigente que solo quería sacar lo mejor de cada persona. Como director de tesis, hacía especial hincapié en la necesidad de hacer artículos que recogiesen buenas ideas que además eran necesarias. En este sentido, siempre se opuso a la fiebre de "publicar por publicar" que parece abundar tanto hoy en día. Además, en el ámbito profesional, siempre creyó que lo académico debía estar ligado a lo empírico, al contexto aplicado. Probablemente, esta fue una de sus grandes virtudes como académico.

Como doctoranda de Julio, siempre gocé de su apoyo profesional y personal en el desarrollo de mi carrera. Junto a Vicente Ponsoda, y como codirector de la Cátedra de Modelos y Aplicaciones psicométricos, hizo posible mi primera oportunidad profesional como psicométra para poder desarrollar mi tesis de manera remunerada. Siempre le estaré agradecida por ello. Para él, era muy importante que un doctorando estuviera financiado para que pudiera centrarse exclusivamente en este trabajo y pudiera tener la oportunidad de desarrollarse de forma debida a nivel académico. Además, una de sus preocupaciones fundamentales siempre ha sido el disponer de un proyecto del Ministerio para poder financiar a su equipo y que pudieramos asistir a congresos. Durante el periodo de desarrollo de mi tesis, yo misma he asistido a varios de los congresos nacionales e internacionales más importantes en el campo de la metodología de las ciencias del comportamiento.

Su último gran proyecto profesional era la organización del XVI Congreso de Metodologia de las Ciencias del Comportamiento y de la Salud, el cual se celebrará el próximo julio de este año, en la Facultad de Psicología de la Universidad Autónoma de Madrid. La organización tiene muy presente la figura y memoria de Julio.

Pero hablar de Julio como profesional implica también hablar de su calidad humana. Porque para él un buen profesional no era solo aquel que era capaz de dar muy buenas clases o de publicar muchos artículos al año. Para él un buen profesional tenía que tener una serie de aptitudes y cualidades personales, en definitiva, una serie de valores, que le dieran un sentido a lo meramente laboral. Para Julio, era muy importante ir al lugar de trabajo a diario no solo para trabajar, que es la finalidad primera pensarán muchos probablemente, sino también "para estar disponible por si algún compañero necesita algo de ti", o "por si algún alumno de repente necesita ayuda y sube a tutoría". Julio siempre fomentó la honestidad y el rigor, la rectitud, la constancia y el buen hacer, y la ilusión por encima de todo, entre sus compañeros, doctorandos, y alumnos, y siempre rechazó actitudes competitivas, deshonestas y arrogantes. Él creía que el ambiente de trabajo tenía que ser saludable, y creía por encima de todo en las relaciones humanas, en que uno tenía que ser capaz llevarse bien con todo el mundo. Tanto es así, que algunos le han definido como "el pegamento" que era capaz de unir a todo el mundo.

La mejor manera de demostrarnos esto a todos sin decir nada era predicando con el ejemplo. Si alguien caminaba a su lado en el pasillo, de seguro no duraba más de medio minuto sin que alguna persona parase a Julio para saludarle o comentarle algo. Reflejo de ello después de que que se haya ido fueron las numerosas condolencias que llegaron al correo de sus compañeros procedentes de otros compañeros y amigos de distintas partes de España. Si no lo hubiera vivido por mí misma y me lo hubieran contado, tal vez hubiera sido difícil de creer, o tal vez no, pero el día de su fallecimiento, y también el de su funeral, fueron numerosas las personas que se acercaron a mostrar sus condolencias y apoyo a la familia. Sin exagerar, al nivel de una "estrella de rock", como alguien le describió una vez también.

Han pasado ocho meses desde que Julio no está entre nosotros, pero todos le tenemos presente, o al menos yo le tengo muy presente, en el día a día. Muchas veces, más de las que probablemente soy consciente, me sorprendo pensando: "Julio hubiera pensado esto seguro", o "Seguramente Julio habría hecho esto". A veces también recuerdo muchos momentos personales, anécdotas o comentarios graciosos que él hacía, o incluso pensando bromas que él haría ante situaciones que ocurren en la actualidad. Y aunque es muy fácil reírse porque él tenia

un gran sentido del humor, especialmente para reírse de sí mismo, es inevitable que siga habiendo momentos en los que las lágrimas nos afloren en los ojos a todos. De momento, y probablemente por mucho tiempo, esto va a continuar siendo así, aunque todos le recordamos con mucho cariño.

En lo personal, como digo, me resulta imposible no acordarme de él todos los días por algún motivo, grande o pequeño. Era una persona que me aportaba tanto, no solo en lo profesional sino también en lo personal, que es difícil no notar su ausencia. Siempre tenía algún comentario por hacer, como el de "Para cuando nos envías el artículo" cuando se impacientaba por leer tu trabajo, o el de "Dolores, te vas a conquistar las Américas" antes de marcharme de estancia. Esta fue la última vez en la que pude verle y darle un abrazo. Además, como Julio tenía una gran presencia, y quien le ha conocido lo sabe, el vacío que dejó para mí a mi vuelta a España fue y sigue siendo especialmente notable. Sin duda, estaria muy contento de que al aula más grande de toda la facultad, el Aula Magna, le hayan puesto su nombre: Julio Olea.

Pero más allá de todo lo dicho, está claro que Julio era una persona excepcional, de esas que cuesta encontrar y a las que no se puede olvidar. Por ello, me gustaría que tampoco se olvide y que quede reflejado en estas hojas el papel clave que ha tenido en el desarrollo de esta tesis más allá de lo meramente formal, que no ha sido poco, y de lo que ha significado para mí como mentor y maestro en lo profesional, y también como padre metodológico en lo personal. Porque aunque no pueda estar a mi lado el día de la defensa, aún hoy en día y en el futuro me encantaría saber qué es lo que pensaría, diría, o haría en esa u otra situación, y me seguiré acordando de él los miércoles en la comida psicométrica, o al dar la vuelta psicométrica a la facultad, o cuando escuche cantar al coro de la facultad en Navidad. Y también cuando coja mi tesis entre mis manos, y piense: esta es la última tesis de Julio Olea, y yo he tenido durante mucho tiempo la suerte de ser su doctoranda, la última.

Loli Nieto.

Madrid, 28 de Mayo del 2019

Agradecimientos/Acknowledgements

Esta tesis ha sido posible gracias al apoyo incondicional de muchas personas que han formado parte de ella de una u otra forma a lo largo de estos años.

En primer lugar, quiero dar las gracias a mis directores de tesis Francisco José Abad, Luis Eduardo Garrido, y Julio Olea, por haberme brindado todo su apoyo y confianza durante esta etapa. Me considero una privilegiada al haber podido formarme con tres personas tan extraordinarias en el plano profesional, pero más aún en su calidad humana.

Paco, has sido y serás la motivación que me llevó a iniciar este viaje. Siempre te voy a estar agradecida por haberme enseñado que soy capaz de superar cualquier límite que me proponga en lo profesional, pero sobre todo por haber sido un gran amigo y un apoyo hasta el final y más allá. Sin ti, nada de esto hubiera sido posible.

Luis, siempre has sido un soplo de aire fresco, pero especialmente en la recta final. Tu paciencia, pasión e ilusión por esta profesión me animan a querer ser mejor profesional, pero también a disfrutar de forma especial mi trabajo. Siempre te voy a estar agradecida por tu gran calidad humana y por tu ayuda en mi estancia en Charlottesville.

Julio, has sido el motor que ha mantenido mi ilusión en el camino. Aunque siga echando en falta a veces escuchar unas palabras tuyas, siempre tengo presentes tu vitalidad y tu rectitud en mi día a día. Gracias por haber apostado por mí y por haberme dado la mejor de las oportunidades en el Laboratorio 17.

Son tantos los momentos que guardo con cariño con vosotros tres a lo largo de este viaje que nunca los voy a olvidar. Espero honraros en lo profesional y en lo personal durante mi carrera, pero sobre todo seguir siendo fiel a mis valores tal y como vosotros me habéis enseñado a través de vuestro ejemplo.

Quisiera dar las gracias también a Vicente Ponsoda por la oportunidad en la Cátedra de Modelos y Aplicaciones Psicométricos. A Carmen García, por ser un ejemplo de profesionalidad y por su gran apoyo durante todo este tiempo. Ambos sois para mí un ejemplo de humildad, y junto a Paco, Julio, y mis compañeros del Laboratorio 17, habéis sido, sois y seguiréis siendo para mí una segunda familia (¡muy psicométrica!).

Durante este tiempo, también me he encontrado con el apoyo constante de varios profesores del área de Metodología como son Carmen Ximénez, Javier Revuelta, Ricardo Olmos y Juan Botella. A todos, os quiero agradecer vuestra confianza a través las distintas oportunidades que me habéis brindado en el camino desde que comencé mis estudios de Máster.

I would like to especially thank Hudson Golino for giving me a great opportunity and one of the best experiences in my life through my research stay with him at the University of Virginia. I will never forget everything you did for me, and I am looking forward to welcoming you (and also Mariana and little Ceci and Bea!) to Spain with open arms and a big smile just like you did the first time we met in Charlottesville.

I also wanted to thank the people of Charlottesville who made my research stay an even better experience, especially Ninoska Abreu for all her help, and Erik Ruzek and his family, and Dingjing Shi, for the moments we spent together.

Quisiera dar las gracias de manera especial a mis amigos Miguel Sorrel (¡yo sí te pongo el apellido por si aún quedan dudas de quién eres!), David Moreno, Blanca Requero, David Santos, y José Ángel Martínez por haber sido el mejor grupo de apoyo en esta etapa tan complicada.También a todos aquellos compañeros que en algún momento han sido parte del Laboratorio 17 y con los que he tenido el placer de coincidir.

Gracias también a la gente que en el día a día hace de la facultad un lugar mejor: Salu, Ana, Robert, Raúl, Mari y los demás empleados de la limpieza, y también Edu y su equipo en la cafetería.

A Kamal Djelloul, María Mallo, Bea Burgos, Esther Salido, Borja Solovera, Azahara Navarro y Cari Prieto. ¡Gracias por confiar en mí y por seguir estando ahí al final del camino!

Por supuesto, gracias a mis padres, Loli y Pedro, por todo el amor que me habéis dado durante este tiempo, y a mi hermana María Jesús, por haber estado siempre presente a pesar de la distancia. A Larry, por todo el amor y apoyo incondicional que me has mostrado siempre en las noches en vela. A mis tíos, Soledad y Jesús, por haber creído siempre en mí, y a mis primos Pedro Jesús, María y Soledad, por haber sido siempre un apoyo para mí. A mis abuelos María, Tito, Pedro y Luciana, por creer siempre que podría llegar más lejos. Os quiero con todo mi corazón.

A Marisol, Leopoldo, Marta, Esther, y Sara, porque siempre habeís creido que llegaría al final del camino, y al final, ¡llegué!

Y para ti, Ana, mi persona favorita, no tengo palabras que puedan expresar toda la gratitud que siento. Más allá de esta etapa, pero especialmente durante estos años, has sido siempre mi compañera incondicional, mi mayor apoyo, mi refugio, mi ilusión y mi orgullo. Sin ti, todo esto no hubiera merecido la pena y no hubiera sido posible llegar hasta el final del camino. Gracias por tanto.

Abstract

Personality traits remain a primary focus of study in many psychological areas. Notwithstanding the advances achieved with the consolidation of the Big Five model as a common framework of study, personality assessment still presents some limitations that need to be addressed. First, traditional paper-and-pencil questionnaires are quite long for modern evaluation settings where several instruments are administered or testing time is very limited. Second, although some attempts have been made to measure personality more efficiently through computerized adaptive testing (CAT), they have completely ignored the hierarchical nature of domains and facets of personality traits. Third, most personality research and assessment relies on self-report measures, which is well known are sensitive to the influence of item wording effects that can distort research results. Accordingly, this dissertation sought to address these limitations by means of three studies. Study 1 presents the process of construction and calibration of a wide pool to measure the Big Five facets. Results from a post-hoc simulation study demonstrated that the adaptive administration of the items produced accurate facet scores using only a third of the total of the items in the pool. Study 2 goes one step further and illustrates the construction of a CAT based on the bifactor model, which allows to approach the study of the Big Five while considering its hierarchical nature. A post-hoc simulation study demonstrated that the CAT based on the bifactor model is more advantageous to assess the Big Five personality traits than other traditional competing approaches. Finally, Study 3 used Monte Carlo methods to evaluate the impact of three types of item wording effects (careless, item verification difficulty, and acquiescence) on person score estimates and other aspects (model fit, factor loadings, and structural validity) in the context of unidimensional fixed-length texts. Two models were evaluated to this end: the random intercept item factor analysis (RIIFA) model and the traditional model with one substantive factor (1F). Results revealed that, although the RIIFA model was consistently superior in terms of model fit to the 1F model, it was not able to better estimate the uncontaminated person scores and other parametes for any type of wording effect than the 1F model. In conclusion, the three studies included in this dissertation provided a series of tools to measure personality traits more efficiently and contributed to the advancement of knowledge in the area of wording effect measurement.

Resumen

Los rasgos de personalidad siguen siendo objeto de interés en diversas áreas de la psicología. A pesar de los avances logrados con la consolidación del modelo de los Cinco grandes como marco de estudio común, la evaluación de la personalidad todavía presenta algunas limitaciones que deben ser abordadas. Primero, los cuestionarios tradicionales de lápiz y papel son bastante largos para los entornos de evaluación modernos, donde se administran varios instrumentos o el tiempo de evaluación es muy limitado. Segundo, aunque se han hecho algunos intentos para medir la personalidad de manera más eficiente mediante test adaptativos informatizados (TAIs), estos han ignorado por completo la naturaleza jerárquica de los dominios y las facetas de los rasgos de personalidad. En tercer lugar, la mayoría de investigaciones y evaluaciones de la personalidad se basan en medidas de autoinforme, las cuales se conoce son sensibles al efecto de la polaridad de los ítems, pudiendo distorsionar los resultados de investigación. En consecuencia, esta tesis trató de abordar estas limitaciones mediante tres estudios. El Estudio 1 aborda el proceso de construcción y calibración de un amplio banco de ítems para medir las facetas de los Cinco Grandes. Los resultados de un estudio de simulación *post-hoc* mostraron que la aplicación adaptativa del banco permitió estimar de manera precisa las facetas utilizando solo un tercio del total de los ítems en el banco. El Estudio 2 va un paso más allá e ilustra la construcción de un TAI basado en el modelo bifactor, que permite abordar el estudio de los Cinco Grandes considerando su naturaleza jerárquica. Un estudio de simulación *post-hoc* demostró que el CAT basado en el modelo bifactor es más ventajoso para evaluar los rasgos de personalidad de los Cinco Grandes que otros enfoques tradicionales alternativos. Finalmente, el Estudio 3 utilizó simulación Monte Carlo para evaluar el impacto de tres tipos de efectos de la polaridad de los ítems (inatención, dificultad de verificación del ítem y aquiescencia) en las estimaciones de puntuación de la persona y otros aspectos (ajuste del modelo, pesos factoriales y validez estructural) en el contexto de los test de longitud fija unidimensionales. Se evaluaron dos modelos para este fin: el modelo de análisis factorial de ítems de intercepto aleatorio (RIIFA en inglés) y el modelo tradicional con un factor sustantivo (1F). Los resultados indicaron que, aunque el modelo RIIFA fue superior en términos de ajuste al modelo 1F de manera consistente, no permitió estimar mejor las puntuaciones no contaminadas ni otros parámetros para cualquier tipo de efecto de la polaridad del ítem frente al modelo 1F. En conclusión, los tres estudios incluidos en esta tesis proporcionaron una serie de herramientas para medir los rasgos de la personalidad de manera más eficiente y contribuyeron al avance del conocimiento en el área de la medición de los efectos de la polaridad de los ítems.

Contents

Chapter 1

General Introduction

1.1 A Brief History of This Dissertation

Personality traits have been a frequent target of study across many fields of psychology for decades. The consolidation of the Big Five model as the dominant paradigm in personality research over the last decades (John, Naumann, & Soto, 2008) has laid much of the foundations of what psychologists understand by "personality" nowadays (McCrae & Costa Jr., 2008). Simultaneously, advances in measurement with item response theory (IRT) have allowed the development of computerized adaptive testing (CAT) as a means to improve the efficiency of traditional testing by only administering items tailored to the ability of the examinee (Wainer, 2000; Weiss, 1985). Under this premise, it seems clear that CAT may be an especially suitable framework to measure something as idiosyncratic as personality. Nevertheless, at the very beginning of this dissertation, the applications of CAT to the measurement of the Big Five were still very limited (e.g., Makransky, Mortensen, & Glas, 2013; Reise & Henson, 2000). More specifically in Spain, it lacked an instrument to adaptively assess the Big Five. Being aware of this fact, the first study of this dissertation, presented in Chapter 2, aimed to develop and calibrate the item pool that would be the basis for the first Spanish CAT to measure the Big Five traits of personality (Nieto et al., 2017).

Contemporarily, theorists and researchers conceive personality traits as hierarchically organized: each of the five broader traits (i.e., domains) subsumes several narrower traits (i.e., facets; Costa & McCrae, 1995; Soto & John, 2009). To this respect, in the last two decades, the application bifactor model has increased importantly in the measurement of constructs with a hierarchical or multifaceted structure such as the Big Five personality traits (Reise, 2012). In this context, the development of multidimensional CATs based on the bifactor model (MCAT-B) has increased importantly in the last decade to measure multifaceted constructs, mostly in the field of psychopathology, such as depression, anxiety, and schizotypal personality (Gibbons et al., 2008, 2012, 2014; Gibbons, Weiss, Frank, & Kupfer, 2016; Moore, Calkins, Reise, Gur, & Gur, 2018; Sunderland, Batterham, Carragher, Calear, & Slade, 2017; Weiss & Gibbons, 2007; Zheng, Chang, & Chang, 2013). The second study of this thesis (Nieto, Abad, & Olea, 2018) proposes and illustrates for the first time the application of MCAT-B as an optimal framework to provide efficient estimates of the Big Five domains and facets.

Moreover, the presence of wording effects in self-report measures, broadly used in personality research, is a prevalent issue that has concerned researchers for decades (e.g., Cronbach, 1946, 1950; Jackson & Messick, 1958). Recent studies continue stating that wording effects are ubiquitous in psychological measures such as the Big Five personality dimensions and thus cannot be ignored (Biderman et al. 2011). In despite of the attempts previously mentioned to measure the Big Five adaptively, research concerning the control of wording effects in personality CATs is nonexistent. I conducted a pilot study that aimed to control wording effects in the CATs previously developed by modeling them during the item pool calibration phase. The main research question was whether it was possible to obtain uncontaminated person scores estimates in CAT through the modeling of wording effects. To do so, the random intercept item factor analysis (RIIFA) model (Maydeu & Coffman, 2006) was applied during the item pool calibration phase. This model is probably the most popular to

model wording effects because, in addition to the simplicity to apply it, it has proven to be superior over competing models (Savalei & Falk, 2014). Surprisingly, preliminary results revealed that modeling the wording effects did not lead to a better recovery of the uncontaminated person scores in CAT. Then, I realized that systematic studies evaluating the recovery of person scores in the presence of different wording effects, and more specifically using the RIIFA model, were nonexistent. Thus, the third study arises from the inherent need to better understand how wording effects influences the estimation of parameters, and more specifically, the person scores, when applying the RIIFA model. This study is conducted in the context of fixed length tests because it is necessary to fully understand the impact of modeling or ignoring wording effects during the item pool calibration phase, that is, previously to apply a CAT.

The three studies presented in this dissertation have been developed under the supervision of Dr. Francisco José Abad, Dr. Luis Eduardo Garrido, and Dr. Julio Olea, and they have been partially supported jointly by three research projects and a grant award:

1. Spanish Ministry of Economy and Competitiveness project: "Computerized adaptive testing based on new psychometric models" (PSI2013-44300-P) [*Studies 1* and *2*]

2. Spanish Ministry of Economy and Competitiveness project: "Multidimensional Computerized Adaptive Tests: Improving calibration and item selection algorithms" (PSI2017-85022-P) [*Studies 2* and *3*]

3. UAM-IIC Chair « Psychometric Models and Applications» [*Studies 1*, *2*, and *3*]

4. Young Researcher Grant Award 2017, Asociación Española de Metodología de las Ciencias del Comportamiento y de la Salud (AEMCCO) [Pilot study related to *Study 3*]

The rest of the dissertation is organized as follows. The remaining sections of Chapter 1 provides a general background on the Big Five model, MCAT-B, and wording effects, and

presents the goals of the dissertation. Chapter 2, Chapter 3, and Chapter 4, are devoted to the three studies conducted. Specifically, Chapter 2 presents the development and calibration of an item pool to measure the Big Five personality facets adaptively in the Spanish context. Chapter 3 illustrates the development of the first MCAT-B to measure the Big Five domains and facets efficiently, and then the MCAT-B is compared in terms of performance with other traditional competing approaches. Chapter 4 systematically evaluates the recovery of person score and other parameter estimates obtained with the RIIFA model in presence of different wording effects (carelessness, item verification difficulty, acquiescence). Chapter 5 provides a general discussion of the results obtained and a summary of the main contributions, as well as limitations and future research lines. Chapter 6 contains the Spanish version of the general discussion. Finally, the list of publications derived from Chapters 1, 5 and 6 is presented, and the list of contributed work until the completion of this dissertation and the published versions of the studies presented in the main text can be found in Appendices A to C.

1.2 The Big Five Model: Taxonomy and Limitations of Traditional Assessment

Over the past decades, the five-factor or Big Five model of personality traits has been consolidated as the dominant paradigm in personality research, laying down much of the foundations of what psychologists understand by "personality" nowadays (McCrae & Costa Jr., 2008). The growth of a common, meaningful language has allowed the emergence and integration of numerous important research findings (Connelly & Ones, 2010), something that is reflected in the increasing number of publications per year (John, Naumann, & Soto, 2008). The Big Five model assumes a hierarchical multifaceted structure with five broad personality traits (e.g., Extraversion), each one containing six narrower traits (e.g., Gregariousness). The broad traits are often referred to as "domains", whereas the narrower trait are often termed "facets".

The most common research interests based on the Big Five model have focused on predictive power of personality regarding other psychological variables mainly related to health (e.g., Cauffman, Kimonis, Dmitrieva, & Monahan, 2009), performance (e.g., Wolfe & Johnson, 1995; Kappe & van der Flier, 2010), and interpersonal relationships (e.g., Koutsos, Wertheim, & Kornblum, 2008), sex and cultural differences in personality traits (e.g., Schmitt, Realo, Voracek, & Allik, 2008), development of personality (e.g., Srivastava, John, Gosling, & Potter, 2003), and validation of questionnaires to measure personality (e.g., Goldberg, 1999; Rammstedt & John, 2007). In relation to this last point, many paper-and-pencil inventories (i.e., fixed-length tests) based on the Big Five taxonomy have been developed. Two of the most used questionnaires in research and applied settings are the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992) and the International Personality Item Pool Representation of the NEO PI-R (IPIP-NEO; Goldberg, 1999). Despite their popularity, previous literature has pointed out some drawbacks that should be considered when administering these questionnaires in contexts of large-scale evaluation (e.g., educational guidance or personnel selection processes) and in measuring patient-reported outcomes measures:

1. These personality inventories are usually very long (Soto & John, 2009) because they are based on the 30 Big Five facets and thus contain many items to assess each facet (e.g., the NEO-PI-R has a total of 240 items, that is, eight per facet, and the IPIP-NEO has 300 items, that is, 10 per facet). Consequently, its usage can produce inefficient and time-consuming individual assessments, and is not recommended in short-time applications or evaluation settings where various questionnaires need to be applied (Rammstedt & John, 2007).

2. Short versions of these inventories have been developed, but this is not the best solution to optimize the accuracy of the measure: they have been designed to assess the

broad domains, thereby ignoring the individual facet scores and even excluding some facets. Consequently, they are less accurate, have less convergent validity, and only partially retain the original facet structure in comparison to their parent scales (Gignac, Bates, & Jang, 2007; Johnson, 2014; McCrae & Costa, 2007). Some examples are the NEO Five-Factor Inventory-3 (NEO-FFI-3; McCrae & Costa Jr, 2007), a 60-item version of the NEO-PI-3 (McCrae, Costa Jr, & Martin, 2005), and the shorter versions of the IPIP-NEO with 20, 50, 60, 100, and 120 items (e.g., Johnson, 2014; Maples, Guan, Carter, & Miller, 2014).

3. Some of these questionnaires (e.g., the different versions of IPIP-NEO) are used mainly in research contexts. Although this has obviously fueled the advancement of personality research, the content and score of the items are in the public domain, which discourages their use in applied settings where psychological evaluation has important consequences for people (e.g., high-stakes contexts).

1.3 Assessing Personality with Computerized Adaptive Testing (CAT)

Advances in measurement with item response theory (IRT) have allowed the development of computerized adaptive testing (CAT). A CAT is a computer-based measure in which each respondent is presented items specifically tailored to his or her individual trait level, which is evaluated and updated according to previous responses. The main advantage of CATs relative to traditional fixed-length tests is that they improve testing efficiency by administering fewer items (Wainer, 2000; Reise & Henson, 2000).

The main core of a CAT is a wide pool containing the items that will be presented to examinees, and that have been previously calibrated. Calibrating the item pool means to estimate the person and item parameters by applying an Item Response Theory (IRT) model, so that they are known and available to provide information about the next item to select during the CAT. Current psychometric literature recommends various analyses that should be

performed as part of the calibration process such as testing the unidimensionality of the constructs assessed and the item fit. Readers are referred to Bjorner, Chang, Thissen, and Reeve (2007), Cook et al. (2007), Revicki, Chen, and Tucker (2015), and Thissen, Reeve, Bjorner, and Chang (2007) for a detailed revision and/or illustration on main methodological guidelines for developing item pool and CATs. Furthermore, the study presented in Chapter 2 describes the entire process followed to develop an item pool to measure the facets of the Big Five model in the Spanish context (Nieto et al., 2017).

The application of CATs to measure personality has increased over the last two decades (e.g., Forbey, & Ben-Porath, 2007; Rudick, Yam, & Simms, 2013; Simms, & Clark, 2005). Regarding the Big Five model, pioneer attempts have been conducted to primarily measure the facets with CAT based on unidimensional IRT (UIRT) models (Reise & Henson, 2000) and multidimensional IRT (MIRT) models (Makransky, Mortensen, & Glas, 2013). Although both approaches may also be used to measure the domains separately (UIRT) or while considering their intercorrelations (MIRT), none of them allow modeling simultaneously the hierarchical structure of the Big Five domains and their facets. Recently, the bifactor model has re-emerged as an alternative to account for this type of construct-relevant multidimensionality of psychological measures in several fields (Reise, 2012), included personality (e.g., Abad, Sorrel, García, & Aluja, 2018; Chen, Hayes, Carver, Laurenceau, & Zhang, 2012). The study presented in Chapter 3 illustrates for the first time the development of a CAT based on the bifactor model to assess the Big Five and its advantages over traditional competing approaches.

1.3.1 The Logic of CAT

The application of a typical CAT requires the programming of an adaptive algorithm in which four main components must be specified: (1) a starting rule, (2) a scoring method, (3) an item selection criterion, and (4) a stopping rule.

1. *Starting rule*. When the CAT initializes, it is common to administer an item with a moderate location parameter, *b* (i.e., between -0.5 and 0.5). The underlying logic is that if it can be assumed that the population under study is normally distributed on the construct being measured, then it is reasonable to apply an item that is informative for a person with average trait level (θ) on such construct (Bjorner, Chang, Thissen, & Reeve, 2007; Embretson & Reise, 2000). Other criteria can be specified to select the first item. For example, if some prior information is available regarding the examinee's trait level, then such information might be used to select an item with a *b* parameter that is optimal for that person.
2. *Scoring method*. Two scoring strategies are employed to obtain the person's trait level estimate (**θ**): Maximum Likelihood (ML) estimation (e.g., Fisher information) and Bayesian estimation (i.e., *maximum a posteriori* and *estimated a posteriori* methods; e.g., Embretson & Reise, 2000; Wainer, 2000).
3. *Item selection criterion*. Corresponding to the previous scoring strategies, there are two main procedures to select the next item to administer: (a) ML methods, which consist in selecting the next item that provides the most psychometric information at the examinee's current θ, and (b) Bayesian methods, that involve selecting the item that minimizes the examinee's expected posterior standard deviation, or in other words, the item that makes the examinee's standard error the smallest (Embretson & Reise, 2000; Wainer, 2000).
4. *Stopping rule*. It is common to set two types of stopping rules based on (a) the number of items administered (fixed length CAT) or (b) a specific level of precision required (variable length CAT; e.g., Cook et al., 2007).

In a CAT, the adaptive algorithm starts by selecting an item according to the starting rule defined. For example, the CAT can select for all the respondents the item that maximizes

the Fisher information at $\theta = 0$. Second, according to a respondent's answer, the θ estimate is obtained via a scoring procedure (e.g., maximum a posteriori). Third, the next item is selected according to the defined criteria. For example, it can be specified that the item that maximizes the Fisher information evaluated at the current examinee θ is selected. These steps are repeated until the algorithm reaches the stopping rule, and then the respondent final θ is estimated according to the scoring method previously defined (i.e., maximum a posteriori).

1.3.2 CAT Based on Unidimensional Models (UCAT)

In the simplest modality, CATs based on UIRT models (UCAT) has been applied to assess a single facet at a time. A unidimensional model assumes that there is a single primary latent dimension that explains the correlations between items. The studies of Reise and Henson (2000) found that measuring the Big Five facets with 4-item UCATs provided very precise estimates (average correlation, $\bar{r}$, between UCAT and pool facet scores was higher than .90 and equal to .95 in each study, respectively). Despite this, UIRT models and therefore UCATs are inefficient to measure the Big Five personality traits due to two reasons: (a) they do not consider the intercorrelations between facets of the same domain, and (b) they do not allow to model the hierarchical structure defined by a domain and its facets.

1.3.3 Multidimensional Traditional CAT (MCAT)

As an alternative to UCATs, MIRT based on the correlated-factors model and therefore MCAT based on such a model, allows studying the correlations between several factors to obtain efficient test scores. Makransky et al. (2013) compared the performance of short fixed-length versions, UCAT, and MCAT in measuring the NEO-PI-R facets of each domain. The MCAT approach resulted to be the more efficient procedure, especially when the facets of a domain were highly correlated. Despite MIRT, and thus MCAT, allow modeling the intercorrelations between facets of the same domain, they are still limited approaches to measure the Big Five model because they do not allow representing each domain and its facets simultaneously.

1.3.4 MCAT Based on the Bifactor Model (MCAT-B)

In recent years, the bifactor model has been "rediscovered" (Reise, 2012) for its application in the measurement of defined constructs with specific related facets. A bifactor model specifies that the covariances between a set of items is explained by the effect of a general factor, which accounts for the common variance among all the items, and the effect of several specific factors, each of which explains additional common variance (i.e., residual variance) shared by a group of items that is not accounted by the general factor. Regarding the Big Five model, a bifactor model can be applied to measure each of the domains (i.e., the general factor) and their corresponding facets (i.e., the specific factors).

The interest in the bifactor models has increased dramatically, growing the number of applications in several areas of psychology such as personality (e.g., Abad et al., 2018; Chen et al., 2012) or intelligence (e.g., Abad, Sorrel, Román, & Colom, 2016; Gignac & Watkings, 2013), and also the studies comparing the bifactor with competing models such as the second-order model (e.g., Chen, West, & Sousa, 2006; Gignac, 2016; Reise, 2012). In the last decade, the development of adaptive algorithms based on the bifactor model has increased importantly. Specifically, pioneer MCAT-Bs have been mainly built to measure multifaceted constructs in the field of psychopathology such as depression, anxiety, and schizotypal personality (Gibbons et al., 2008, 2012, 2014; Gibbons et al., 2016; Moore et al., 2018; Sunderland et al., 2017; Weiss & Gibbons, 2007; Zheng et al., 2013). These studies have shown great savings in the number of administered items when using MCAT-B, leading to great savings in testing time. For example, Gibbons and colleagues (2012, 2014) developed adaptive algorithms to estimate the levels of depression and anxiety through MCAT-B. With only 12 items on average (the item pools had 400 items approximately), and with an average of just over two minutes per respondent, the adaptive procedure produced highly accurate estimates (associated standard errors below .3) and high rates of specificity and sensitivity. Besides, Weiss and Gibbons

(2007) developed an MCAT-B to measure the mood or anxiety disorder, showing a reduction of items and time over 90% compared to the application of the complete pool. Haley et al. (2009) also developed a MCAT-B to measure global physical health in children with cerebral palsy (their parents responded the items), allows to estimate the construct with an average of 10 to 15 items. Nieto, Abad, & Olea (2018) illustrated for the first time the development of a MCAT-B to measure the Big Five personality traits. They concluded that, unlike other competing approaches, the use of MCAT-Bs constitutes a preferential framework to measure the Big Five model because it allows assessing each general domains while representing the multidimensionality due to the specificity of its facets. Results from this study are shown in detail in Chapter 3.

1.4 An Introduction to Wording Effects

Most self-report scales in Psychology often include both positively worded (PW) and negatively worded (NW) items to measure a given construct. PW items are intended to measure the presence of a construct with positive valence (e.g., Extraversion), whereas NW items measure the presence of a construct with negative valence (e.g., Introversion; Kam & Meyer, 2015a; Kam, 2016, 2018). Frequently, both PW and NW items measure the two poles of the same construct. For example, a personality scale may include several PW items to measure Extraversion (e.g., I make friends easily) and some NW items to measure Introversion (e.g., I prefer to be alone). However, when both types of items are combined, respondents may manifest differential response styles to PW and NW items. This phenomenon is known as item wording effect and consists of logically inconsistent answers to PW and NW items that tap into similar (but polar opposite) content (Kam & Meyer, 2015a, Kam, 2016).

A prevalent concern regarding the use of self-report measures in psychological measurement is the ubiquity of wording effects and its potential influence on examinees responses (Biderman et al. 2011; Paulhus, 1991; Paulhus, & Vazire, 2005). An extensive body

of research has demonstrated that wording effects may impact the psychometric properties of scales, deteriorating model fit (Abad et al., 2018; Danner, Aichholzer, & Rammstedt, 2015; Woods, 2006), spuriously increasing the dimensionality due to the emergence of separate factors for PW and NW items (Barnette, 2000; Marsh, 1996; Rodebaugh et al., 2004; Schmitt & Stults, 1985), reducing the reliability of measures (Roszkowski & Soven, 2010; Schriesheim, Eisenbach, & Hill, 1991), inflating of suppressing the structural relationships (Kam & Meyer, 2015b; Kam, Zhou, Zhang, & Ho, 2012), and distorting the factor loading structures (Navarro-González, Lorenzo-Seva, & Vigil-Colet, 2016; Savalei & Falk, 2014; Zhang, Noor, & Savalei, 2016).

In this regard, one area of interest that has received less attention is the estimation of the person scores in the presence of wording effects. A plausible reason is that most studies investigating wording effects are conducted using data collected in applied settings (e.g., Wetzel & Carstensen, 2017), making it impossible to know the uncontaminated true score of the respondents. It is common to observe not only in personality assessment but in other psychological areas how different respondents use the response scale idiosyncratically but in a consistent fashion. The presence of wording effects (e.g., acquiescence) can be a potential reason for the emergence of these individual differences in scale usage (Maydeu-Olivares & Coffman, 2006). If such differences in the use of the scale are ignored, it might produce that two individuals with the same true score on a given construct present different score estimates. Similarly, two respondents with different true scores that use the response scale differently might present equal observed or estimated scores (Austin, Deary, & Egan, 2006; Wetzel & Carstensen, 2017).

On the other hand, there are very few studies that have systematically evaluated the effects of item wording and these are very limited in the types of wording effects or parameter estimates studied. For example, Schmitt and Stults (1985) and Woods (2006) studied

exclusively the impact of carelessness on the spurious increase of dimensionality and the deterioration of model fit, respectively. Both studies reached similar conclusions: with only 10% of careless respondents a spurious second dimension emerged (Schmitt & Stults, 1985), and a two-factor model was thus preferred over the true unidimensional solution (Woods, 2006). Grønhaug and Heide (1992) simulated acquiescent responses to Likert type items and found that inconsistent responses might distort results from regression and factor analysis. They simulated medium sample sizes (500 respondents) and short tests (10 items). There are even fewer studies that have examined specifically the recovery of person scores and they also have some limitations. Plieninger (2016) evaluated the impact of acquiescence on reliability, validity, and scale scores estimates, and found that its effect was greater in unbalanced scales with fewer NW items. However, he just focused on small sample sizes (200 respondents) and short tests (10 items). Chapter 4 of this dissertation extends previous findings by examining the impact of three wording effects (carelessness, item verification difficulty, and acquiescence) on several parameter estimates, including the recovery of uncontaminated person scores. The results of this study are shown in detail in Chapter 4.

1.4.1 Types of Wording Effects

In this section, a brief conceptualization of the wording effects studied in this dissertation is presented.

Carelessness. This wording effect refers to a pattern of responding in which respondents do not pay attention to item content. Several terms have been used in prior research to refer to this wording effect such as random responding (Meade & Craig, 2012), noncontingent responding (Baumgartner & Steenkamp, 2001), inattentiveness (Johnson, 2005), or insufficient effort responding (Huang, Curran, Keeney, Poposki & DeShon, 2012). In turn, the concept of carelessness has been broadly used to refer to different random and nonrandom response patterns such as fully or partially random responding, using the same

response category (i.e., straight-line responding) or response sequence, or skipping items (e.g., Swain et al., 2008; Johnson, 2005; Meade & Craig, 2012). Some studies have suggested the existence of a systematic (non-random) type of carelessness in which respondents may answer according to expectations that he or she has formed about what is being measured based on the questionnaire instructions or the content of the initial items (Schmitt & Stults, 1985; Woods, 2006). Authors suggesting this variant of carelessness usually associate it to misresponding to NW items.

Item verification difficulty. Swain et al. (2008) conceptualized item verification difficulty as a type of inconsistent responding that occurs when the respondent belief about the construct being measured (i.e., his or her true trait level) mismatches the content of the item being evaluated. For example, for a person who belief that he or she is extroverted (i.e., has a high trait level in Extraversion) will be easier to verify the statement "I am extroverted" than to reject the statement "I am introverted". As it will be explained later, the difficulty to verify the item content will depend on several factors.

Acquiescence. Acquiescence is without any doubt the most popular wording effect in literature. It is conceptualized as the tendency to respond to items using agree categories (i.e., the positive side of the scale) irrespective of their content (e.g., Paulhus, 1991; Weijters, Baumgartner, & Schillewaert, 2013; Wetzel et al., 2016).

1.4.2 Cognitive Processes Underlying Wording Effects: The Response Process Model

Swain et al. (2008) and Weijters and Baumgartner (2012) conceptualized the three wording effects previously described in terms of their underlying cognitive processes. To do so, they used the response process model developed in survey research literature (Tourangeau, Rips, & Rasinski 2000), which consists of four major steps: (a) *comprehension* (attending to the item and interpreting it), (b) *retrieval* (retrieving a relevant belief previously formed from

long-term memory or transferring to working memory of information used to construct a new belief), (c) *judgement* (integrating the information retrieved previously and comparing it to the item representation), and (d) *response* (representing the answer onto the given scale and producing a response).

First, as the systematic carelessness previously described refers to a type of respondent that does not pay enough attention to item content, these examinees would have problems at the initial step of the response process model: they would not complete the comprehension phase satisfactorily (nor the subsequent retrieval and judgement phases) and consequently they would not process item content (Swain et al., 2008; Weijters & Baumgartner, 2012; Weijters et al., 2013).

Item verification difficulty would emerge during the judgement phase of the response process model. Swain et al. (2008) suggested that the item verification process can be explained according to the constituent-comparison model (Carpenter & Just, 1975). This model postulates that a respondent's difficulty to verify an item, and thus the probability of misresponding it, will depend on the complexity in comparing his/her own belief or true trait level on the construct being measured to the item content. This difficulty will depend on the number of cognitive operations that the respondent has to perform to compare his or her belief with item content. And this in turn depends on whether item content is on the same pole (i.e., is truth) or on the opposite pole (i.e., is false) relative to the respondent belief (i.e., true trait level), and whether the item is affirmed or negated. For example, a person whose belief is that he or she is extroverted (i.e., has a high trait level in Extraversion) will have increasing difficulty in responding the following items: "I am extroverted" (true affirmation), "I am introverted" (false affirmation), "I am not extroverted" (false negation), and "I am not introverted" (true negation). Swain et al. (2008) corroborated these predictions through a series of experiments.

Finally, acquiescence will influence the response phase (Swain et al., 2008; Weijters & Baumgartner, 2012; Weijters et al., 2013), which is the final step of the response process model (Tourangeau et al., 2000). Knowles and Condon (1999) suggested that the dual-process model of understanding (Gilbert, 1991) can explain the cognitive process underlying this wording effect. Such model consists of two phases: first, in the comprehension phase, the respondent automatically accepts the item content, and second, in the reconsideration phase, he or she can reevaluate it to decide whether to reject it or continue accepting it. This second step implies an effort for the participant, so it can be omitted depending on his or her ability and motivation. If this occurs, a respondent will automatically agree to all items, irrespective if they are PW or NW (Swain et al., 2008; Weijters & Baumgartner, 2012).

Although some respondents affected by different wording effects may show similar response patterns, it should be noted that they both will present differences during the response process. For example, some types of acquiescent respondents that systematically use the highest response category might resemble some types of careless respondents with a straight-line responding pattern, and vice versa. However, a careless respondent will overlook item content (the problem arises at the initial comprehension phase) whereas an acquiescent one will pay attention to it (the problem occurs at the final response phase; Kam & Meyer, 2015a; Weijters et al., 2013).

1.4.3 Measuring Wording Effects

Traditionally, measurement experts have recommended the construction of balanced scales (i.e., with equal number of PW and NW items; e.g., Nunnally, 1967; Paulhus, 1991) to control for the influence of wording effects such as acquiescence. However, this strategy itself will be only valid if one is interested in computing the scale mean, but it will be useless if a researcher aims to perform some analysis based on the covariance matrix and wording effects influence respondents differently (Savalei & Falk, 2014).

On the other hand, prior studies examining the effects of acquiescence often use measures based on the endorsement of polar opposite items (e.g., Hinz, Michalski, Schwarz, & Herzberg, 2007; Rammstedt & Farmer, 2013) or many items with heterogeneous content (Baumgartner & Steenkamp, 2001; Kam & Zhou, 2015; Weijters et al., 2013). However, some of these measures (i.e., those not based on heterogeneous items) may also reflect other wording effects such as carelessness (Weijters et al., 2013; Kam & Meyer, 2015a), leading to erroneous conclusions about the influence of acquiescence. On the other hand, most research investigating carelessness has mainly focused on the detection of careless respondents through the use of different methods such as Mahalanobis distance or indices based on the number of consecutive items answered with the same response option (for a detailed review, see Curran, 2016). In this regard, instructed response items have been shown to be able to detect careless respondents and to distinguish them from acquiescent ones (Kam & Meyer, 2015a).

To the best of my knowledge, there are no studies that have examined the effects of item verification difficulty on parameter estimates. However, it is likely that the origins of the study of item verification difficulty relies on prior research suggesting the relationship between wording effects and reading ability. In that regard, some classic studies are those of Marsh (1986, 1996), who found that wording effects (in this case associated to NW items) were weaker for more verbally able students.

Another widespread strategy to control for wording effects is the estimation of models that include one or several wording method factors (e.g., Billiet & McClendon, 2000; Marsh, 1986, 1996; Yang et al., 2018). Among the different models defined in prior literature, the random intercept item factor analyses (RIIFA) model has become very popular in the last years. This is probably due to its simplicity to be implemented, and because prior studies have shown that wording effects can be successfully modeled with this model as evidenced by the improvements in model fit in comparison to models that only contain substantive factors (e.g.,

Abad et al., 2018; Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006; Yang et al., 2018).

1.5 The RIIFA Model

Maydeu-Olivares and Coffman (2006) introduced the RIIFA model as an extension of the common factor model that allows for the explicit modeling of consistent individual differences in the use of the response scale. In the common factor model, the response of participant *j* to item *i* (y_{ij}) can be written as:

$$y_{ij} = \mu_i + \boldsymbol{\lambda}_i' \boldsymbol{f}_j + e_{ij}, \tag{1.1}$$

where μ_i is the intercept for item *i*, $\boldsymbol{\lambda}_i$ is the vector of factor loadings for item *i*, $\boldsymbol{f}_j$ is the vector of substantive factor scores for participant *j*, and e_{ij} is the error term for participant *j* on item *i*. Assuming that the mean of the common factors and the error terms is zero, and that the error terms are uncorrelated with each other and with the common factors, the covariance matrix implied by this model ($\boldsymbol{\Sigma}_y$) is expressed as:

$$\sum\nolimits_y = \boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta} \tag{1.2}$$

where *lambda* ($\boldsymbol{\Lambda}$) is a $k \times m$ matrix of factor loadings for k variables and m common factors, *psi* ($\boldsymbol{\Psi}$) is a $m \times m$ covariance matrix of the common factors, and *theta* ($\boldsymbol{\Theta}$) is a $k \times k$ covariance matrix of the error terms.

In the RIIFA model the intercept (γ_{ij}) is decomposed into a fixed part (μ_i) common to all respondents but differing across items, and a random part (ζ_j) common to all items but differing across respondents:

$$y_{ij} = \gamma_{ij} + \boldsymbol{\lambda}_i' \boldsymbol{f}_j + e_{ij} \qquad \gamma_{ij} = \mu_i + \zeta_j \tag{1.3}$$

$$y_{ij} = \mu_i + \zeta_j + \boldsymbol{\lambda}_i' \boldsymbol{f}_j + e_{ij} \tag{1.4}$$

If in addition to the previous assumptions of the common factor model it is assumed that the term ζ_j is standardized and that it is uncorrelated with the error terms and with the common factors, the covariance structure implied by the RIIFA model can be written as:

$$\sum_{y} = \mathbf{1}\omega\mathbf{1}' + \mathbf{\Lambda}\,\mathbf{\Psi}\mathbf{\Lambda}' + \mathbf{\Theta} \tag{1.5}$$

where ω is the variance of ζ_j across all respondents.

In the RIIFA model, the parameter to be estimated is ω and not the random intercept for each examinee. To do so, it is only necessary to define an additional wording method factor in the common factor model in which all the unstandardized factor loadings are fixed to 1 (if items are not reverse coded) and ω is left free to be estimated.

Savalei and Falk (2014) have systematically evaluated the performance of the RIIFA model in estimating item parameters when respondents make an idiosyncratic use of response scale in the context of unidimensional scales. They found that the RIIFA model was superior to competing approaches (including the "do nothing" approach) and robust to the violation of its assumption of equal wording factor loadings across items.

1.6 Goals of the Current Dissertation

All things considered, the main goal of this dissertation is to contribute to the improvement of personality assessment from a methodological approach. Specifically, this dissertation has been structured as a compendium of publications with three studies that were developed relying on both the analysis of real data and the use of Monte Carlos methods. On the one hand, the first two studies are oriented towards addressing the main limitations of traditional fixed-length tests used to measure personality. In this sense, these studies aimed to develop a new measure (a CAT) to assess the Big Five personality traits more efficiently in the Spanish context. On the other hand, a third study aims to advance the area of wording effects. This study sought to understand the impact of different wording effects. In addition, it aims to provide researchers with applied guidelines to avoid misconceptions in the interpretation of

results based on the use of self-report measures affected by wording effects. The specific goals of the three studies are presented below.

1.6.1 Study 1: Calibrating a New Item Pool to Adaptively Assess the Big Five

The main purpose of *Study 1* was to develop an item pool to constitute the basis for the first Spanish CAT to measure the FFM facets efficiently. In addition, this study aimed to test the performance of separate CATs to measure the Big Five model, and more specifically the facets, more efficiently. As result of the analyses performed in this study, an item pool with good psychometric properties to measure the Big Five facets was obtained.

1.6.2 Study 2: Assessing the Big Five with Bifactor Computerized Adaptive Testing

Previous to this dissertation, UCATs and MCATs based on correlated traits were tested to adaptively assess the Big Five model. In addition, short fixed-length versions are a widespread method to assess the Big Five personality traits. However, all these methods had not be compared previously. In addition, these measures ignored the hierarchical structure nowadays sustained by contemporary experts in personality psychology. Thus, the main purpose of *Study 2* was to assess whether a MCAT-B can provide more efficient estimates of the Big Five personality traits than three other competing approaches: a short scale, UCAT, and MCAT with correlated factors. In addition, this study sought to test whether benefits of applying MCAT-B depend on the degree of multidimensionality of the measured Big Five traits.

1.6.3 Study 3: Does Modeling Wording Effects Help Recover Uncontaminated Person Scores?

Although the motivating goal of this study was to advance in the knowledge of the impact of wording effects on person score estimates, it is also common in model evaluation that researchers also pay attention to other model results. Besides, previous research has shown that the RIIFA model can successfully model wording effects compared to models that only

contain substantive factors. Thus, the main goal of *Study 3* was to assess the performance of the RIIFA model to estimate different types of parameters (model fit indexes, factor loadings, person scores, and the relationship with a criterion variable) in the presence of three wording effects previously defined in literature: carelessness, item verification difficulty, and acquiescence.

Chapter 2

Calibrating a New Item Pool to Adaptively Assess the Big Five

Abstract

Background: Even though the Five Factor Model (FFM) has been the dominant paradigm in personality research for the past two decades, very few studies have measured the FFM adaptively. Thus, the purpose of this research was the building of a new item pool to develop a computerized adaptive test (CAT) for personality assessment. *Method*: A pool of 480 items that measured the FFM facets was developed and applied to 826 participants. Facets were calibrated separately and item selection was performed attending to the preservation of unidimensionality of each facet. Then, a post-hoc simulation study was carried out to test the performance of separate CATs to measure the facets. *Results*: The final item pool was composed of 360 items with good psychometric properties. Findings reveal that a CAT administration of four items per facet (total length of 120 items) provides accurate facets scores, while maintaining the factor structure of the FFM. *Conclusions*: An item pool with good psychometric properties was obtained and a CAT simulation study demonstrated that the FFM facets could be measured with precision using a third of the items in the pool.

This chapter contains the accepted version of the following manuscript:
Nieto, M. D., Abad, F. J., Hernández-Camacho, A., Garrido, L. E., Barrada, J. R., Aguado, D., & Olea, J. (2017). Calibrating a new item pool to adaptively assess the Big Five. *Psicothema, 29,* 390–395. http://dx.doi.org/10.7334/psicothema2016.391
The published version of the manuscript is presented in Appendix B.

2.1 Introduction

Over the past 25 years the Five Factor Model (FFM) of personality traits (also called 'Big Five') has been established as the dominant paradigm in personality research, exceeding 300 publications per year (John, Naumann & Soto, 2008). The FFM assumes a multifaceted structure with five broad personality traits (i.e., domains) each one containing several narrower traits (i.e., facets).

Although in personality research there is a debate about the measurement of facets versus domains, many studies have shown that narrow measures contribute to the prediction of several outcomes in various contexts (e.g., Ashton, Paunonen, & Lee, 2014). Thus, most personality tests developed to measure the FFM are based on facets. This is the case for the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992) and the International Personality Item Pool Representation of the NEO PI-R (IPIP-NEO; Goldberg, 1999).

Because the FFM contain many facets, these questionnaires are usually very long (e.g., 240 items for the NEO PI-R), resulting in individual assessments that are oftentimes time consuming and inefficient. As a counter measure, short versions of such scales have been proposed but these have been designed to assess the broad domains, thereby ignoring the individual facet scores and even excluding facets. For example, the NEO Five-Factor Inventory-3 (NEO-FFI-3; McCrae & Costa Jr., 2007) is a version of the NEO PI-R with 60 items taken from 28 of the 30 facet scales. Another characteristic of some personality tests like the IPIP-NEO is that the items are placed in the public domain. Although this has given rise to great advances in personality research, its use could not be recommended in evaluation contexts where examinees must not know the item content prior to the administration.

Advances in measurement with item response theory (IRT) have allowed the application of computerized adaptive testing (CAT) as an alternative to traditional tests in a variety of contexts, including the study of personality. Pioneer attempts have been carried out

recently to measure the Big Five adaptively. Two studies have performed real-data simulations using responses to the NEO-PI-R items. First, Reise & Henson (2000) found that administrating separate CATs for evaluating the FFM facets provided accurate estimates with half of the NEO PI-R items. More recently, Makransky, Mortensen, and Glas (2012) applied separate multidimensional CATs in order to measure the facets on each domain and obtained increases in the reliability of the facet scores. Also, the Tailored Adaptive Personality Assessment System (TAPAS) is a CAT used to measure the FFM in military settings in the United States (e.g., Stark, Chernyshenko, Drasgow, & White, 2012). Recently in Spain, Pedrosa, Suárez-Álvarez, García-Cueto, and Muñiz (2016) developed a CAT to assess specific personality traits of enterprising personality in young people.

The main core of a CAT is the wide pool of items that is calibrated with an IRT model (i.e., the person and item parameters are known). In the Reise and Henson (2000) and Makransky et al. (2012) studies the items of the NEO-PI-R were calibrated, thereby creating an item pool. However, because a number of phases are involved in an item pool construction, the current psychometric literature recommends other rigorous analyses that should be performed before starting the calibration such as testing the unidimensionality of the constructs and the fit at the item level (e.g., Revicki, Chen, & Tucker, 2015).

In view of all the above, we present in this study the development of an item pool to constitute the basis for the first Spanish CAT to measure the FFM facets efficiently. To do so, we identify four major steps: (a) develop items of each facet and obtain evidence for content validity, (b) calibrate each facet separately, checking the unidimensionality assumption and IRT fit, (c) test the performance of separate facet CATs, and (d) obtain evidences for internal structure and convergent validity. Thus, the specific purposes of this study were (a) to design, calibrate, and validate a new item pool based on the FFM and (b) to study the performance of CATs to measure the FFM facets more efficiently.

2.2 Method

2.2.1 Participants

A sample of 871 psychology undergraduate students participated voluntarily in the study. The sampling was intentional. Preliminary analyses revealed that a low percentage of the participants (45 respondents, 5.16% of the initial sample) presented careless, invalid or atypical responses according to multiple criteria described in the data analysis section and were consequently excluded. The final sample was composed of 826 individuals aged 17 to 50 years (M = 20.06, SD = 3.73), of which 696 were female (70.91%). For some analyses, the whole sample was randomly divided into two datasets with equal size (n = 413), one for applying exploratory statistical analysis (model-derivation sample) and the other one for validating statistical results (validation sample). The University Research Ethics Committee granted approval for the present study. The full anonymized data set is available from the authors upon request.

2.2.2 Instruments

Personality item pool. According to the traditional descriptions of the FFM facets, four independent experts in personality assessment and psychometrics developed an initial pool of 480 items (16 per facet) in Spanish language. The recommendations for item pool building were followed (e.g., Revicki et al., 2015). Then, each expert reviewed the item content of the whole pool and redundant statements were excluded and replaced by new ones. The statements were administered using a five-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). A Spanish philologist revised the items and corrected grammar, spelling and style errors. Table 2.1 shows facets 1 to 6 for each domain.

Directed questions scale. A scale of 12 Likert-type items (1 = strongly disagree, 5 = strongly agree) directing participants to give specific responses (e.g., “If you are reading this

question, please mark 'Disagree'") was applied to measure inattention. Scale scores were obtained by summing the correct responses.

NEO-FFI-3. The NEO-FFI-3 inventory, a 60-item version of the NEO-PI-3 (McCrae, Costa Jr, & Martin, 2005) to measure the FFM domains, was included to obtain evidences for convergent validity of the new item pool. The NEO-PI-3 is a revision of the NEO PI-R. Due to there are no Spanish versions of the NEO-PI-3 and the NEO-FFI-3 questionnaires, 59 of the 60 items of the NEO-FFI-3 were selected from the Spanish version of the NEO-PI-R (Cordero, Pamos, & Seisdedos, 2008). The remaining item was translated from the English version of the NEO-FFI-3.

Table 2.1. Five Factor Model: Domains and Facets

	Domain				
Facet	Neuroticism	Extraversion	Openness	Agreeableness	Conscientious-ness
1	Anxiety	Warmth	Fantasy	Trust	Competence
2	Angry/hostility	Gregariousness	Aesthetics	Straightforward-ness	Order
3	Depression	Assertiveness	Feelings	Altruism	Dutifulness
4	Self-consciousness	Activity	Actions	Compliance	Achievement striving
5	Impulsiveness	Excitement seeking	Ideas	Modesty	Self-discipline
6	Vulnerability	Positive emotions	Values	Tender-mindedness	Deliberation

2.2.3 Procedure

The items from the personality item pool, the Directed Questions scale and the NEO-FFI-3 were used to create two booklets that were administered in two sessions in a counterbalanced order. Participants completed the items within an official system of data collection in a faculty of Psychology whose purpose is the participation of students in research projects in exchange for academic compensation.

2.2.4 Data Analysis

Evidence for content validity. Evidence for content validity of the personality item pool was obtained. Thirty-six experts in personality research and psychometrics were asked to select the facet to which each item belonged. Each expert evaluated the items from two domains. The level of congruence between the experts for each item was measured as the percentage of classification agreement for its most chosen facet. After excluding the responses from experts with low reliability (i.e., percentage of congruence lower than 70% in at least one domain), items with less than 50% of classification in their corresponding theoretical facet were removed from the pool.

Personality item pool IRT calibration. Psychometric properties of the pool were analyzed by fitting the unidimensional graded IRT response model (Samejima, 1969) to each subset of items measuring the same facet. First, some indexes were examined in order to screen out data for careless, invalid or atypical responses (i.e., score below 9 points on the Directed questions scale, double responses in more than three items, more than 10 missing values on the personality items, outliers regarding the number of consecutive identical responses).

For each facet, the unidimensionality assumption was tested on the model-derivation sample by applying parallel analysis (PA) and the unidimensional factor model with the polychoric correlation matrix and the robust unweighted least squares (ULSMV) estimator. If unidimensionality was not tenable according to PA or some variables had very low factor loadings, items were iteratively removed until the unidimensionality assumption was met and all the items had factor loadings larger than .2. For purposes of achieving unidimensionality, the highest residual correlation was identified and the item with the smaller loading in this pair was deleted. At the end of the iterative process, PA and the comparative fit index (CFI) were used, as recommended in Garrido, Abad, & Ponsoda (2016) to assess the unidimensionality of

facets in the cross-validation sample. The conventional cutoff values for the CFI, are .90 or greater for acceptable fit, and .95 or greater for good fit (Hu & Bentler, 1999).

The selected subset of unidimensional items of each facet was calibrated separately according to the graded IRT response model using the Metropolis-Hastings Robbins-Monro algorithm (MHRM; Cai, 2010a, 2010b) on the whole sample. Item fit was tested on the sample with complete response patterns using the polytomous variant of the $S - X^2$ index (Orlando & Thissen, 2000) with the Benjamini-Hochberg adjustment to control Type I error (Benjamini & Hochberg, 1995). Finally, the IRT maximum a posteriori (MAP; Embretson & Reise, 2000) pool facet scores and the standard errors (*SE*s), indicating the precision of trait estimates (θ), were obtained for each individual in each facet. IRT marginal reliabilities for pool facet scores were also obtained (Brown & Croudace, 2015; p. 314).

Performance of the CAT. A post hoc simulation study was carried out to analyze the performance of the CATs in measuring the FFM facets. We simulated a separate CAT for each facet using the item responses obtained from the respondents. Since omissions are not allowed in CATs, the response vectors were completed using item and respondent estimated parameters obtained in the previous calibration step. The CAT algorithm started by selecting the item that maximized the Fisher information at $\theta = 0$ for all the respondents. Then, attending to a respondent answer, the MAP θ estimate was obtained. The next item selected was the one that maximized the Fisher information evaluated at the θ estimate. These steps were repeated until the algorithm stopped when four items were administered. Then, the final CAT facet score was estimated using the MAP method.

Different criteria were used to analyze the precision of the CATs. For each facet, the correlation between the CAT and the pool scores were obtained. We also obtained the empirical reliability and the median of the *SE* across examinees for each CAT score.

Evidence for internal structure and convergent validity for pool and CAT facet scores. First, evidence based on the factorial structure of the pool facet scores was obtained. PA with Pearson correlations was used to verify that the suggested number of factors was five as expected (one factor per personality domain). Next, we applied exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009) with the maximum likelihood estimator. Unlike exploratory factor analysis, ESEM models can include both exploratory and confirmatory methods (e.g., correlated error terms). Using the model-derivation sample, we defined five correlated ESEM factors corresponding to the five domains. The Oblimin rotation method was used. Since modification indices suggested some correlated residuals, a new model including them was tested using the cross-validation dataset. Again, PA and the CFI were used for model evaluation. Additionally, the same ESEM factor model was used to test the internal structure of CAT facet scores. Factor congruence coefficients were obtained to study the similarities of the factorial structure obtained with pool and CAT scores.

Following the previous step, pool and CAT domain scores were obtained as an average of the correspondent six facet scores. Composite reliabilities for domain scores were estimated from the ESEM models as the squared correlation between the domain trait score and the corresponding latent factor (Raykov, 1997). Finally, evidence for convergent validity was obtained by computing the correlations between the CAT and the pool domain scores with the NEO-FFI-3 raw scores.

All the analyses were performed M*plus* 7 (Muthén & Muthén, 1998-2012) and the R packages *psych* (Revelle, 2016), *mirt* (Chalmers, 2012), and *mirtCAT* (Chalmers, 2016).

2.3 Results

Evidence for content validity. Two experts out of 36 were excluded by their low percentage of congruence (below 70%) in the Extraversion domain. After excluding these experts, the average percentages of congruence by domain were 84% for Neuroticism, 86% for

Extraversion, 93% for Openness, 89% for Agreeableness, and 86% for Conscientiousness. Twenty-five items out of 480 were removed from the item pool by their low percentage of classification in the theoretical facet (less than the 50%). After excluding these items, the average percentages of classification accuracy by domain were 89% for Neuroticism, 87% for Extraversion, 94% for Openness, 90% for Agreeableness, and 89% for Conscientiousness.

Personality item pool IRT calibration. Out of 871 participants 45 were excluded from the sample of analysis because they presented careless, invalid or atypical responses. Missing data rate for item nonresponse was very low with a maximum value of 2%.

Out of 455 items 95 were removed in order to preserve the unidimensionality of each facet. The largest number of excluded items in one facet was 7 (i.e., in the Assertiveness, Straightforwardness, and Dutifulness facets). For the retained items, the unidimensionality assumption was always tenable according to PA. The unidimensional solution showed acceptable fit according to the CFI, which was equal or above .90 in 80% of the cases and equal to or higher than .85 in the remaining facets (except for Tender-mindedness, CFI = .62). PA indicated that the 67% of the facets were unidimensional. In the remaining facets, PA suggested a two-factor solution (except for Excitement seeking that PA indicated three factors). In this cases, the scree test revealed that the second empirical eigenvalue was barely greater than the random eigenvalue. All the item factor loadings on the unidimensional solutions were statistically significant ($p < .05$), with average loadings ranging from .45 to .73.

Within the framework of the IRT, only 4 items out of 360 were identified as misfitting to the graded response model according to the $S - X^2$ index. The *a-parameter* of the items showed adequate positive values ranging from 0.35 to 3.86 ($\bar{a}$ = 1.51), with 23% of them being highly discriminative (i.e., $a > 2$).

Figure 2.1 illustrates the information and *SE* for each θ pool facet scores. For θ between –3 and 3, the *SE*s for almost all the facets, except Compliance and Dutifulness, were lower than

.5, which is approximately equivalent to a reliability coefficient of .75. This indicates that the items provide good information across the different traits levels of each facet, except for the two facets mentioned. Regarding marginal reliability, all facet scores presented values equal to or above .72. Average reliabilities for pool facet scores within a domain were .89, .90, .88, .85 and .86 for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness, respectively.

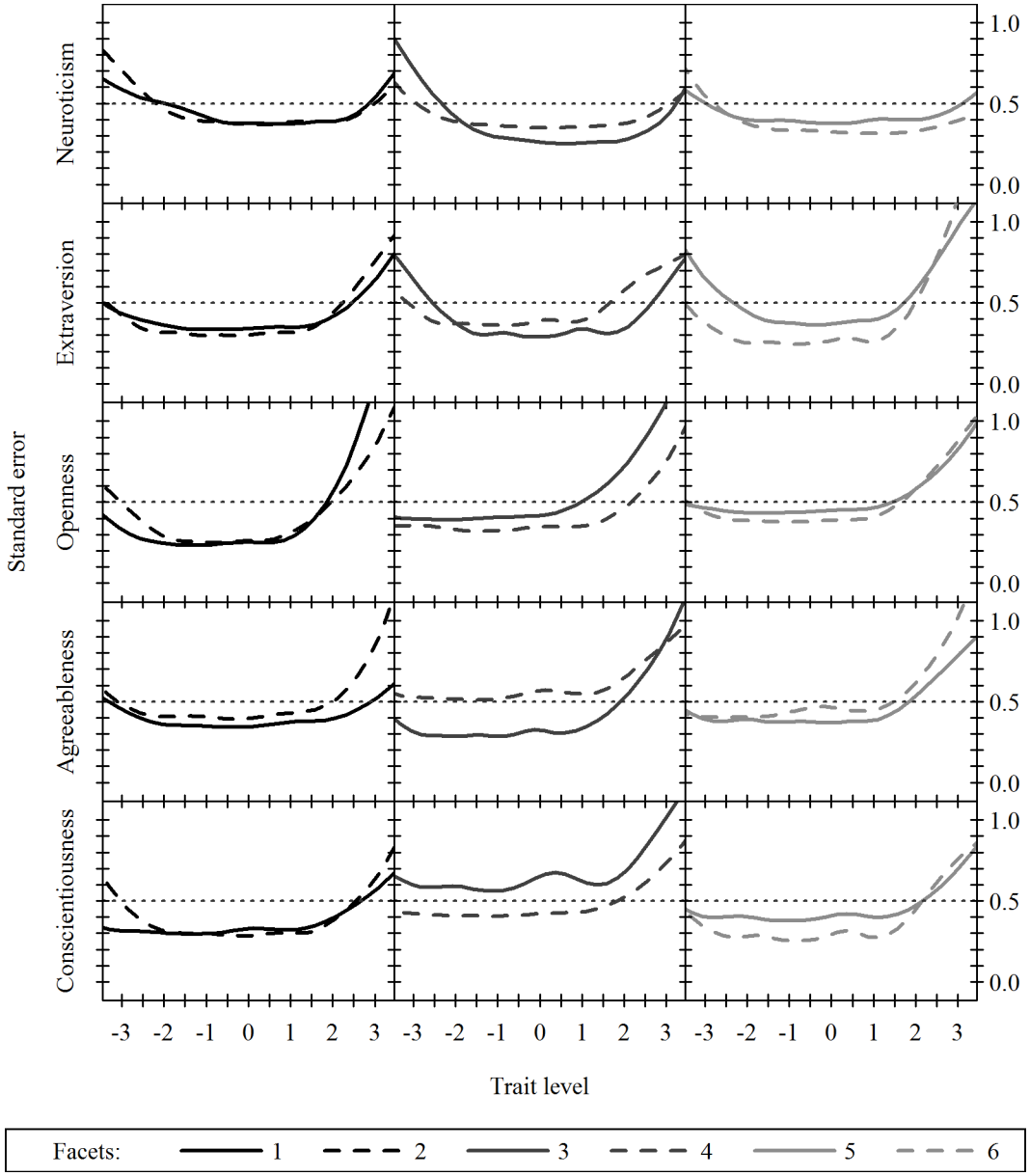

Figure 2.1. Standard error (*SE*) across the trait level for the facets of each domain of the FFM. *SE* equal to .50 is indicated with a dotted line. The facets 1 to 6 of each domain are specified in Table 1.

Performance of the CAT. Correlations between each CAT facet scores and pool facet scores were high for all the facets with values ranging from .92 to .98 ($\bar{r}$ = .95). For most facets, the median of the participants' *SE* was lower than .4. Only Ideas ($Mdn_{SE} = .41$), Compliance ($Mdn_{SE} = .48$), Tender-mindedness ($Mdn_{SE} = .41$), and Dutifulness ($Mdn_{SE} = .53$) presented higher values. Regarding marginal reliability, most facet scores presented values equal or above .7, except the Dutifulness facet with a value of .68. Average reliabilities for pool facet scores within a domain were .82, .86, .81, .79 and .79 for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness, respectively.

Evidence for internal structure and convergent validity for pool and CAT facet scores. As expected, PA based on the analysis of the pool facet scores suggested five factors. Thus, a five-factor exploratory model was first specified for the ESEM analyses in the model-derivation sample. This model was then modified adding six correlated residuals according to modification indexes above 40. Correlated residuals were theoretically meaningful (e.g., a negative correlation between Deliberation and Impulsiveness) and were replicated in the validation sample in which the modified model fit was acceptable: CFI was .91 and PA indicated a five-factor solution.

In the final modified model, almost all the facet scores loaded higher and significantly on its respective domain factor. These loadings were medium-high sized with values above .40 (M = .61). Only the Social anxiety and Deliberation facets presented values below .40 (.35, and .31, respectively). Regarding cross-loadings, most of them were on the Extraversion (Depression: –.33, Social anxiety: –.63, Impulsiveness: .45, Actions: .39, Trust: .35, and Deliberation: –.43), Agreeableness (Angry/hostility: –.35, Feelings: .37, Dutifulness: .35, and Deliberation: .30), and Openness (Emotions seeking: .38; Order: –.36) domains. Also Activity and Competence facets cross-loaded .33 and –.44 on Conscientiousness and Neuroticism, respectively. Average cross-loading (in absolute value) was low (.14).

The factor correlation matrix showed that Neuroticism correlated negatively with Extraversion (r = –.28; p < .001), and Conscientiousness (r = –.21; p < .001). Additionally, Extraversion also correlated, positively, with Openness (r = .24; p < .001) and Conscientiousness (r = .23; p < .001). Conscientiousness was also correlated with Openness (r = .12; p < .001) and Agreeableness (r = .10; p < .001). The remaining correlations were small ($|r|$ < .06).

When the ESEM was applied to the CAT facet scores, the results were highly similar (i.e., congruence coefficients were .99 for each of the five factors). Composite reliabilities for pool domain scores were acceptable and ranged from .75 (Agreeableness) to .87 (Extraversion). Reliabilities for CAT domain scores were inferior as expected but acceptable and ranged from .70 (Openness) to .86 (Extraversion). According to the Spearman-Brown formula and the pool composite reliabilities, it must be noted that in order to obtain these 24-item length CAT domain score reliabilities, 56 items would be required, in average, in a fixed form.

Finally, correlations between the pool domain scale scores and the NEO-FFI-3 raw scores were good. The Extraversion and Neuroticism domains presented the highest convergent validity values (r = .88 and .86, respectively). In the case of Openness and Agreeableness scales the value was similar (r =.83), and Conscientiousness presented the lowest value (r =.80). Convergent validity for the CAT domain scale scores with the NEO-FFI-3 were only slightly inferior (the largest difference, .02, was for Neuroticism).

2.4 Discussion

Recent studies in personality have investigated the possibility of obtaining accurate personality facet scores with CATs (e.g., Makransky et al., 2012). The purpose of this research was to build a new personality item pool and develop the first Spanish CAT based on the FFM facets. Analyses were performed at the facet-level. This is one of the key aspects of this study

because recent research has shown that facet-level analysis increases the predictive validity of personality scores (Ashton et al., 2014).

In this study a pool of items for personality assessment is provided and efficiently administered with CAT. Although there are several commercial paper-and-pencil tests for assessing the FFM, this might be an important contribution to the evaluation of personality in applied settings where short-time assessments are required and the item content should be unknown to the examinees prior to administration.

Four main steps are distinguished in this study. First, item statements were developed and evidence for content validity was obtained via the evaluation of experts. Second, each facet was calibrated separately according to the Samejima graded response model. Unidimensionality of facets was guaranteed through a strict iterative analysis procedure and almost all the items showed adequate fit to the Samejima graded response model. In terms of precision, the facet scales showed generally good reliability with small *SE* over a wide range of θ. In line with previous studies (e.g., Benet-Martínez & John, 1998) and the NEO PI-R manuals, the facets of the Neuroticism, Extraversion and Openness domains were, on average, the most reliable.

Third, a CAT simulation study revealed that using separate 4-item CATs to assess the facets (i.e., with an administration of 120 items), facet scores are estimated accurately with low SEs in most cases. Finally, internal structures of the pool and the CAT were analyzed obtaining similar results: facets in both instruments measured the narrow traits of their corresponding FFM domains. Some facets loaded on more than one domain (e.g., Angry/hostility was designed to measure a subdomain of Neuroticism and was also an indicator of Agreeableness). This is consistent with previous studies that have shown that an important part of the variance of the facets scales is due to different domains (e.g., Abad, Sorrel, García, & Aluja, in press).

In addition, both the item pool and CAT scores showed good convergent validity with the NEO-FFI-3 questionnaire.

One limitation of the current study is the generalizability of the results to other samples, although the intercorrelations found between the five personality factors are consistent with previous research. For example, Neuroticism correlated negatively with Extraversion and Conscientiousness, and Extraversion also correlated positively with Openness (e.g. Mount, Barrick, Scullen, & Rounds, 2005; Van der Linden, te Nijenhuis, & Bakker, 2010). Furthermore, domains such as Neuroticism and Openness showed lower correlations. However, due to the fact that the sample consisted of psychology undergraduate students, we are aware that the results may not be generalized to other sub-populations (e.g., clinical, workforce).

Recent research has suggested that multidimensional IRT models and multidimensional CATs may increase the precision of personality trait scores (e.g., Makransky et al., 2012). In this regard, future research with the presently developed item pool should be oriented toward the application of multidimensional models in the calibration and adaptive administration phases.

References

Abad, F. J., Sorrel, M. A., García, L. F., & Aluja, A. (In press). Modeling general, specific, and method variance in personality measures. Results for ZKA-PQ and NEO-PI-R. *Assessment*. doi: 10.1177/1073191116667547

Ashton, M. C., Paunonen, S. V., & Lee, K. (2014). On the validity of narrow and broad personality traits: A response to Salgado, Moscoso, and Berges (2013). *Personality and Individual Differences, 56*, 24-28. doi: 10.1016/j.paid.2013.08.019

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397-438. doi: 10.1080/10705510903008204

Benet-Martínez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait–multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75*, 729–750. doi: 10.1037/0022-3514.75.3.729

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*, 289-300. doi: 10.2307/2346101

Brown, A. & Croudace, T. J. (2015). *Scoring and estimating score precision using multidimensional IRT*. In Reise, S. P. & Revicki, D. A. (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (pp. 334-363). New York, NY: Routledge.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hasings Robbins-Monro algorithm. *Psychometrika, 75*, 33-57. doi: 10.1007/s11336-009-9136-x

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*, 307-335. doi: 10.3102/1076998609353115

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1-29. doi: 10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*, 1-39. doi: 10.18637/jss.v071.i05

Cordero, A., Pamos, A., & Seisdedos, N. (2008). NEO PI-R, Inventario de Personalidad NEO Revisado [Revised NEO Personality Inventory]. *Madrid: TEA Ediciones.*

Costa, P., & McCrae, R. R. (1992). *NEO PI-R manual professional.* Odessa, FL: Psychological Assessment Resources, Inc.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, *21*, 93. doi: 10.1037/met0000064

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe, 7*, 7-28. Tilburg, The Netherlands: Tilburg University Press.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi: 10.1080/10705519909540118

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. John, R. Robins, & L. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). New York, Guilford.

Makransky, G., Mortensen, E. L., & Glas, C. A. (2012). Improving personality facet scores with multidimensional computer adaptive testing: an illustration with the NEO PI-R. *Assessment, 20*, 3-13. doi: 10.1177/1073191112437756

McCrae, R. R., Costa, Jr, P. T., & Martin, T. A. (2005). The NEO–PI–3: A more readable revised NEO personality inventory. *Journal of Personality Assessment*, *84*, 261-270. doi: 10.1207/s15327752jpa8403_05

McCrae, R. R., & Costa Jr., P. T. (2007). Brief versions of the NEO-PI-3. *Journal of Individual Differences, 28*, 116-128. doi: 10.1027/1614-0001.28.3.116

Mount, M. K., Barrick, M. R., Scullen, S. M., & Rounds, J. (2005). Higher-order dimensions of the big five personality traits and the big six vocational interest types. *Personnel Psychology, 58*, 447-478. doi: 10.1111/j.1744-6570.2005.00468.x

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide (7th ed.).* Los Angeles, CA: Muthén & Muthén.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50-64. doi: 10.1177/01466216000241003

Pedrosa, I., Suárez-Álvarez, J., García-Cueto, E., & Muñiz, J. (2016). A computerized adaptive test for enterprising personality assessment in youth. *Psicothema, 28,* 471-478. doi: 10.7334/psicothema2016.68

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21,* 173-184. doi: 10.1177/01466216970212006

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7,* 347-364. doi: 10.1177/107319110000700404

Revelle, W. (2016) Procedures for personality and psychological research. Evanston, Il: Northwestern University.

Revicki, D. A., Chen, W. H., & Tucker, C. (2015). Developing item banks for patient-reported health outcomes. In Reise, S. P. & Revicki, D. A. (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 334-363). New York, NY: Routledge.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. *Psychometrika, 34(Suppl. 1),* 1-97. doi: 10.1007/BF02290599

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods, 15,* 463-487. doi: 10.1177/1094428112444611

Van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44,* 315-327. doi: 10.1016/j.jrp.2010.03.003

Chapter 3

Assessing the Big Five with Bifactor Computerized Adaptive Testing

Abstract

Multidimensional computerized adaptive testing based on the bifactor model (MCAT-B) can provide efficient assessments of multifaceted constructs. In this study, MCAT-B was compared with a short fixed-length scale and computerized adaptive testing based on unidimensional (UCAT) and multidimensional (correlated-factors) models (MCAT) to measure the Big Five model of personality. The sample comprised 826 respondents who completed a pool with 360 personality items measuring the Big Five domains and facets. The dimensionality of the Big Five domains was also tested. With only 12 items per domain, the MCAT and MCAT-B procedures were more efficient to assess highly multidimensional constructs (e.g., Agreeableness), whereas no differences were found with UCAT and the short scale with traits that were essentially unidimensional (e.g., Extraversion). Furthermore, the study showed that MCAT and MCAT-B provide better content-balance of the pool because, for each Big Five domain, items from all the facets are selected in similar proportions.

This chapter contains the accepted version of the following manuscript:
Nieto, M. D., Abad, F. J., & Olea, J. (2018). Assessing the Big Five With Bifactor Computerized Adaptive Testing. *Psychological Assessment, 30*, 1678–1690. http://dx.doi.org/10.1037/pas0000631
The published version of the manuscript is presented in Appendix C.

3.1 Introduction

The Big Five model of personality traits has been established as the dominant paradigm in personality research, exceeding 300 publications per year (John, Naumann & Soto, 2008). The Big Five model assumes a hierarchical multifaceted structure with five broad personality traits (i.e., domains) each one containing six narrower traits (i.e., facets). Although in personality research, there is a debate about the measurement of facets versus domains (Salgado et al., 2014), many studies have shown that narrow measures contribute to the prediction of several outcomes in various contexts (e.g., Ashton, Paunonen, & Lee, 2014; McAbee, Oswald, & Connelly, 2014; O'Connor & Paunonen, 2007). Thus, major personality inventories based on the 30 Big Five facets are usually very long due to the fact that they contain many items to assess each facet. This is the case for the NEO Personality Inventory-3 (NEO-PI-3; McCrae, Costa, & Martin, 2005) with a total of 240 items (i.e., 8 items per facet) and the International Personality Item Pool Representation of the NEO PI-R (IPIP-NEO; Goldberg, 1999) with 300 items (i.e., 10 per facet). Consequently, these questionnaires lead to individual assessments that are inefficient and time-consuming and are not recommended in short-time applications or evaluation contexts where various questionnaires need to be applied. As a countermeasure, short versions of such scales have been developed. For example, the NEO Five-Factor Inventory-3 (NEO-FFI-3; McCrae & Costa Jr., 2007) is a 60-item version of the NEO-PI-3 (McCrae et al., 2005), although there are others. Likewise, brief versions of the IPIP-NEO have been developed, such as the IPIP-NEO-120 (e.g., Johnson, 2014; Maples, Guan, Carter, & Miller, 2014). However, these shortened questionnaires have been designed to assess the broad domains, thereby ignoring the individual facet scores and even excluding some facets. Consequently, they are less accurate than the original versions, have less convergent validity with their parent scales as the number of items decreases, and only partially

retain the original facet structure (Gignac, Bates, & Jang, 2007; Johnson, 2014; McCrae & Costa Jr., 2007).

Advances in measurement with item response theory (IRT) have allowed the application of computerized adaptive testing (CAT), improving the efficiency of traditional testing by only administering items tailored to the ability of the examinee. In personality research, pioneer attempts have been conducted to measure the Big Five adaptively using CAT based on unidimensional (UCAT; Nieto et al., 2017; Reise & Henson, 2000) and multidimensional (correlated-factor) models (MCAT; Makransky, Mortensen, & Glas, 2012). These studies have shown high gains in efficiency over the administration of the complete test. On another hand, the interest in the bifactor model has increased dramatically due to its effectiveness to represent multifaceted constructs such as the Big Five personality traits (Reise, 2012). Indeed, Abad, Sorrel, García, and Aluja (2016) have endorsed the potential of MCAT based on the bifactor model (MCAT-B) for this purpose. However, the bifactor model has not been applied so far to adaptively assess the Big Five.

In this study, we propose that applying MCAT based on the bifactor model (MCAT-B) can provide efficient estimates of the Big Five domains and facets. In addition, we suggest that MCAT-B can provide more accurate estimates than other approaches (e.g., short scales, UCAT, MCAT). The article is structured as follows: First, we will outline some issues about the evaluation of personality with CAT. We then provide a short background about recent applications of MCAT-B. Next, we will describe the procedure followed in this study to calibrate items according to the bifactor model in order to later apply MCAT-B. Then, we will evaluate the efficiency of score estimates on the Big Five using four different procedures for each domain (a short scale, UCAT, MCAT, and MCAT-B). Finally, we will address practical implications of adaptively assessing the Big Five personality traits. The analyses proposed in this study will be carried out using a new item pool designed to evaluate the Big Five model.

3.2 Assessing Personality with Computerized Adaptive Testing

The application of CAT to measure personality has increased over the last decades (e.g., Forbey, & Ben-Porath, 2007; Rudick, Yam, & Simms, 2013; Simms, & Clark, 2005). Specifically, in the case of the Big Five model, CAT developments have been based on the unidimensional IRT (UIRT) model to assess a single facet at a time (see Figure 3.1, model A). The UIRT model assumes that there is a single primary latent dimension which explains the correlations between items. In this regard, Reise and Henson (2000) found that evaluating the facets of the NEO PI-R separately with 4 items through UCAT provided accurate trait estimates in comparison with the complete 8-item facet scales (r > .90). Similar results were obtained by Nieto et al. (2017). They applied UCAT to assess each facet with 4 items of a new item pool based on the Big Five model and found an average correlation of $\bar{r}$ =.95 between UCAT and pool facet scores. However, the application of separate UIRT models and therefore UCAT does not allow considering the intercorrelations between facets of the same domain. Consequently, the fact of ignoring such information makes UIRT inefficient to represent the Big Five personality traits.

On the other hand, MIRT based on the correlated-factors model and, by extension, MCAT based on such a model, allows studying the correlations between several factors to obtain efficient test scores (see Figure 3.1, Model B). Makransky et al. (2012) demonstrated that the application of MIRT improved the precision and efficiency of the NEO PI-R facets when they were highly correlated. Thus, the facets of Neuroticism, Openness, and Conscientiousness, which showed the highest intercorrelations on average ($\bar{r}$ = .70 for the former, and $\bar{r}$ = .60 for the two last), obtained greater gains in precision. In addition, applying MCAT to model the facets of each domain led to facet scores as least as accurate as UIRT on average, with reductions in test length of 75% for Neuroticism, 63% for Openness, and 50% for Conscientiousness.

Although both UIRT and MIRT approaches have been applied to study the Big Five facets, they do not allow modeling simultaneously multiple hierarchically organized constructs that represent a broad trait (i.e., the domain) and several narrower subdomains (i.e., the facets). The application of the bifactor model has increased dramatically as an alternative to account for this type of construct-relevant multidimensionality of psychological measures in several fields (Reise, 2012).

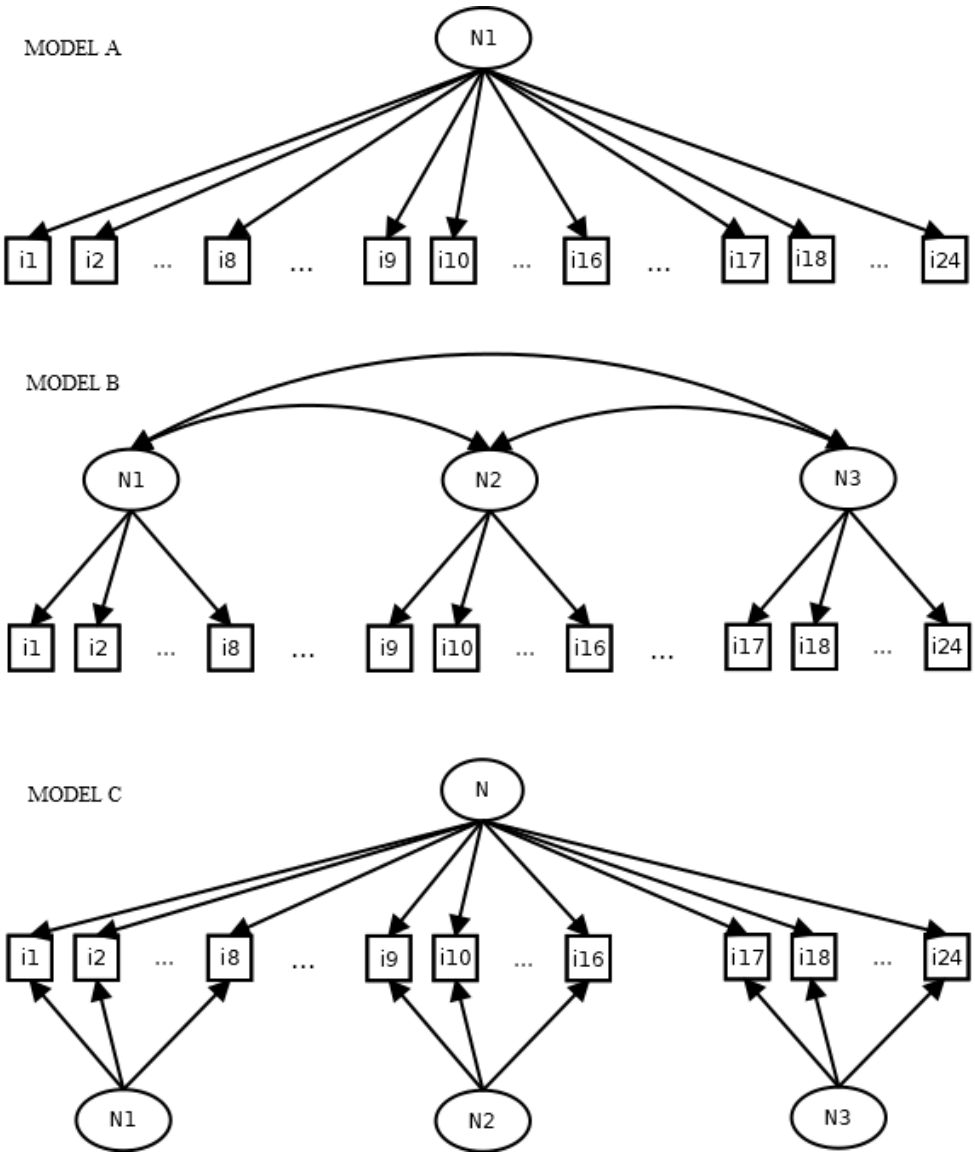

Figure 3.1. Representation of three different models for the Neuroticism (N) domain and three of its facets (N1 = Anxiety, N2 = Hostility, and N3 = Depression). Model A = Unidimensional; Model B: Multidimensional Correlated Traits; Model C: Bifactor. i1,…, i24 represent the items.

3.3 Computerized Adaptive Testing Based on the Bifactor Model

In the bifactor model, each item loads simultaneously on a general factor (i.e., domain) and on one of the several specific factors (i.e., facet) that account for additional common variance between clusters of items that is not explained by the general factor. All the dimensions (i.e., general and specific) are first-order factors that are assumed to be orthogonal. In Figure 3.1, Model C is depicted an example of a bifactor model, with a general factor representing the Neuroticism domain and three specific facets: Anxiety, Hostility, and Depression.

In personality research, several studies have applied the bifactor model to assess the Big Five traits. Chen, Hayes, Carver, Laurenceau, and Zhang (2012) illustrated the use of the bifactor model to test the multifaceted structure of the Extraversion domain of the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992). Abad et al. (2016) employed the bifactor model in order to separate the sources of variance due to the general and specific factors in each of the Big Five traits of the NEO PI-R. In addition, the application of MCAT-B has increased importantly in the last decade, mostly in the field of psychopathology, to measure multifaceted constructs such as depression, anxiety, and schizotypal personality (Gibbons et al., 2008, 2012, 2014, 2016; Moore, Calkins, Reise, Gur, & Gur, in press; Sunderland, Batterham, Carragher, Calear, & Slade, 2017; Weiss & Gibbons, 2007; Zheng, Chang, & Chang, 2013). These studies have shown great savings in the number of administered items when using MCAT-B: Reductions of up to 97% were found when estimating domain scores whereas reductions ranging from 67% to 85% were found when also assessing the specific facets. In these studies, MCAT-B improved measurement precision, with CAT trait estimates being highly correlated with those obtained with the full item pool (i.e., correlations above .90). In addition, important reductions in the time required to complete the evaluations have been reported. For example, Gibbons et al. (2012) found that with a mean of 12 items, an

average of 2.29 minutes was enough to estimate the trait level in the depression severity domain in comparison with the 51.66 minutes required to complete the full 389- item test.

3.4 Proposal for the Current Study

Taking all of the above into account, we propose that applying MCAT-B might provide a more suitable approach to assess the Big Five because its key feature includes modeling simultaneously the variance due to each broad domain and its narrower facets. To the authors' knowledge, the performance of MCAT-B has not been compared with UCAT and MCAT based on correlated traits to assess the Big Five model. In addition, proposed short fixed-length versions of large Big Five inventories neither have been compared to different MCAT procedures. Thus, the main aim of this study is to assess whether a MCAT-B can provide more efficient estimates of the Big Five personality traits than three other competing approaches: a short scale, UCAT, and MCAT with correlated factors. Additionally, we study whether benefits of applying MCAT-B depends on the degree of multidimensionality of the measured Big Five trait: It is expected that the bifactor model will be more advantageous with highly multidimensional traits, whereas the unidimensional approach will be preferred for traits with a strong general factor. Therefore, a secondary goal is to examine whether item responses to the Big Five personality traits are sufficiently unidimensional to apply UIRT methods instead of bifactor and other MIRT models.

3.5 Method

3.5.1 Participants and Procedure

The dataset includes responses from 826 undergraduate psychology students (696 women [70.91%], 175 men [20.09%]) to a pool with 360 personality items to evaluate the Big Five traits. Participants' ages ranged from 17 to 50 years (M = 20.06, SD = 3.73). For some analyses, the whole sample was randomly divided into two datasets with equal size ($n_1 = n_2 =$ 413), one for model-derivation analysis and the other one for cross-validating statistical results.

Participants completed the items in a Psychology Faculty within an official system of data collection whose purpose was the participation of students in research projects in exchange for academic compensation. The University Research Ethics Committee granted approval for the present study.

3.5.2 Instruments

Personality item pool. The pool is composed of 360 items rated on a five-point Likert-type scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*), measuring the Big Five and their facets: Neuroticism (Anxiety, Angry/hostility, Depression, Self-consciousness, Impulsiveness, and Vulnerability), Extraversion (Warmth, Gregariousness, Assertiveness, Activity, Excitement-seeking, and Positive emotions), Openness (Fantasy, Aesthetics, Feelings, Actions, Ideas, and Values), Agreeableness (Trust, Straightforwardness, Altruism, Compliance, Modesty, and Tender-mindedness), and Conscientiousness (Competence, Order, Dutifulness, Achievement striving, Self-discipline, and Deliberation). Statements are written in the Spanish language.

Details of the original validation of the pool are provided in Nieto et al. (2017). The items of each facet were calibrated according to the unidimensional model. Average alpha coefficients for the facets within each domain ranged from .85 (Agreeableness) to .90 (Extraversion). Within the UIRT framework, the standard error (*SE*) for trait levels θ between –3 and 3 were lower than .50 for all the facets except for Compliance (Agreeableness) and Dutifulness (Conscientiousness), which is approximately equivalent to a reliability coefficient of .75. The analysis of the internal structure using pool facet scores revealed that the items were properly designed to measure the Big Five factors of personality. The pool also showed excellent convergent validity with the NEO-FFI-3 scales, with correlations ranging from .80 to .88.

NEO-FFI-3. An external measure, the NEO-FFI-3, was included in order to examine the convergent validity of the item pool calibrated according to the bifactor model. The NEO-FFI-3 is a 60-item version of the NEO-PI-3, which is in turn a revision of the NEO PI-R, that provides measures for the Big Five domains of personality. Due to the fact that Spanish versions of the NEO-PI-3 and the NEO-FFI-3 questionnaires are not available, 59 of the 60 items of the NEO-FFI-3 were selected from the Spanish version of the NEO-PI-R (Cordero, Pamos, & Seisdedos, 2008). The remaining item was translated into Spanish from the English version of the NEO-FFI-3.

3.5.3 Data Analysis

Calibrating each domain separately: application of IRT bifactor model. First, the missing data rate was analyzed at the item level in the whole data set. Then, the model-derivation sample ($n_1 = 413$) was used to estimate separate exploratory bifactor graded response models (Gibbons et al., 2007) for each personality domain: A structure with a general factor representing the domain and as many specific factors as facets was specified. The Metropolis-Hastings Robbins-Monro algorithm (MHRM; Cai, 2010a, 2010b) was used for parameter estimation. The MHRM method allows missing item responses. In order to identify each model, marker items (i.e., those with the highest factor loading on their corresponding facet according to the unidimensional model) were specified to load only on their corresponding specific factor and on the general factor, whereas the remaining items were allowed to load on all the factors. With regard to the non-marker items, minimally informative normal prior distributions N (0, .10) were specified for the slopes of the facets on which they theoretically should not load. Then, items with factor loadings below .20 on the general factor were excluded in an iterative procedure. At the end of this process, facets with less than 5 items were excluded from the analysis.

Subsequently, the cross-validation sample (n_2 = 413) was used to test the model previously estimated for each domain. Five fit indices were obtained for model evaluation: the M_2^* statistic for polytomous data (Cai & Hansen, 2013), the root mean square error of approximation (RMSEA) as calculated from the M_2^* values (Maydeu-Olivares, Cai, & Hernández, 2011), the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the standardized root mean square residual (SRMSR). CFI and TLI values of .90 or greater indicate acceptable fit, and values of .95 or greater represent good fit. For the RMSEA and SRMSR indices, values between .05 and .08 are indicative of acceptable model fit, whereas values below .05 represent good fit (Hu & Bentler, 1999; McDonald & Ho, 2002). Finally, the item parameters of each model were estimated in the whole sample (N = 826).

Degree of essential unidimensionality of the domains. Two bifactor-specific indices were computed: the explained common variance (*ECV*) and the proportion of uncontaminated correlations (*PUC*). The *ECV* (Sijtsma, 2009; Ten Berge & Sočan, 2004) reflects the common variance due to the general factor and can be easily calculated. For example, for a bifactor model with six specific factors (i.e., one per facet):

$$ECV = \frac{\sum \lambda_G^2}{\sum \lambda_G^2 + \sum \lambda_{s_1}^2 + \sum \lambda_{s_2}^2 + \sum \lambda_{s_3}^2 + \sum \lambda_{s_4}^2 + \sum \lambda_{s_5}^2 + \sum \lambda_{s_6}^2} \quad (3.1)$$

where λ_G are the factor loadings of the general factor and λ_{s_1} to λ_{s_6} are the factor loadings of the specific factors s_1 to s_6. High *ECV* values (e.g., above .85 or .90), indicate a strong general factor, so that data can be considered essentially unidimensional and therefore modeled using UIRT without seriously biasing parameter estimates. Values below .70 reflect that data are sufficiently multidimensional and MIRT models should be applied (Quinn, 2014; Stucky & Edelen, 2014).

Reise, Scheines, Widaman, and Haviland (2013) and Bonifay, Reise, Scheines, and Meijer (2015) showed that the *ECV* is related to parameter bias and that the *PUC* is an

important moderator in this relationship. The *PUC* (Bonifay et al., 2015; Reise et al., 2013) indicates the proportion of between-item correlations that, according to the theoretical model, are not affected by the specific factors. For each Big Five domain, the *PUC* was computed according to its theoretical independent cluster structure. For the previous example with six facets, the *PUC* can be calculated as:

$$\frac{J_G \times (J_G - 1) - [J_{s_1} \times (J_{s_1} - 1) + J_{s_2} \times (J_{s_2} - 1) + ... + J_{s_6} \times (J_{s_6} - 1)]}{J_G \times (J_G - 1)} \quad (3.2)$$

where J_G is the total number of items of the domain and J_{s_1} to J_{s_6} are the number of items of the specific factors s_1 to s_6. Following the authors previously mentioned, as the *PUC* increases, the *ECV* becomes less important to determine the extent of parameter bias. In general terms, when the *PUC* is very high (e.g., > .90), even low *ECV* values can yield unbiased parameter estimates (e.g., Reise, 2012). Rodríguez, Reise, and Haviland (2016a) suggested that when both *ECV* and *PUC* are >.70, low parameter bias is found.

In order to quantify the parameter distortion resulting from fitting multidimensional (bifactor) data to a unidimensional model, the relative bias (*RB*) was computed for each item as the difference between the loading on the one-factor model and the general loading on the bifactor model divided by the general factor loading on the bifactor model (Rodríguez, Reise, & Haviland, 2016b). Then, for each domain, the overall *RB* was computed as the average of the individual *RB*s in absolute value for the items of the domain. Values below 10–15% indicate minor bias (Muthén, Kaplan, and Hollis, 1987).

Precision and evidence for convergent validity of pool scores. The alpha coefficient (α) was obtained in order to assess the precision of the domain and facet pool raw scores. Then, within the framework of bifactor MIRT, the multidimensional maximum a posteriori (MAP; Bock & Aitkin, 1981) method was used to obtain the trait estimates (θ) for examinees in the

domains and their facets. The precision of θ estimates was evaluated with the associated standard errors (*SE*s).

In addition, evidence for convergent validity was obtained by computing the Pearson correlation coefficients (*r*) between the pool raw scores on the domains and the NEO-FFI-3 raw scores.

Post-hoc simulation study. A post-hoc simulation study (i.e., drawing simulees' responses from the real data) was carried out to compare, for each of the Big-Five traits, the performance of four procedures: a fixed-length short scale, UCAT, MCAT, and MCAT-B. Since several items were excluded from the initial 360-item pool in the previous calibration step, only the responses to the final 307-item pool were used to simulate the four methods. As omissions are not allowed in CAT, each examinee's response pattern was completed using item and person parameter estimates obtained in the previous calibration step with the bifactor model.

The MCAT-B was based on the bifactor model. The items were adaptively selected according to the D-Optimality criteria (i.e., maximize the determinant of the information matrix for trait estimates; see Seo & Weiss, 2015). For selecting the first item, traits ($\boldsymbol{\theta}$) were initialized to zero and from there on, MAP $\boldsymbol{\theta}$ estimates ($\hat{\boldsymbol{\theta}}$) were computed according to the respondent answers. CAT stopped when 12 items were administered. For each MCAT-B, $\hat{\boldsymbol{\theta}}$ estimates were obtained for one general and several specific factors. It must be noted that, in the bifactor model, the specific factors reflect the residual variance after subtracting the effects of the general domain. That is, they reflect whether the examinee facet score is above or below the expected score after controlling for the general factor (DeMars, 2013). Previous research has suggested that removing domain-level variance may dramatically alter the meaning of the facet-level constructs, so that this residualized facet scores may have an ambiguous meaning (e.g., Simms, Prisciandaro, Krueger, & Goldberg, 2012). For this reason, and for ease of

interpretation, the expected or predicted observed scores on the facets (τ_f) and the domains (τ_d), which reflect the respondent's overall standing on each scale, were obtained. For example, the expected score of a respondent in facet *f* was obtained as the sum of the expected scores on the items measuring it (DeMars, 2013):

$$\tau_f = \sum_{j \in f} \sum_{k=1}^{K} k P_{jk}(\hat{\theta}) \tag{3.3}$$

where *k* runs from 1 to K, the number of response categories, and $P_{jk}(\hat{\theta})$ is the probability for a respondent with a $\hat{\boldsymbol{\theta}}$ estimate of selecting response category *k* of item *j*.

For the UCAT and the MCAT, the same CAT specifications were used, but based on the UIRT and the MIRT models, respectively. Thus, in order to apply these procedures, data were calibrated separately for each domain according to the UIRT (i.e., one general factor for all the items in the domain) and the MIRT models (i.e., one factor per facet). Again, for comparability with the MCAT-B, expected scores were obtained based on $\hat{\boldsymbol{\theta}}$ estimates. Finally, for each domain, a fixed-length short scale was developed with the 12 items with the highest factor loadings on the UIRT model. Expected scores were again obtained, based on $\hat{\boldsymbol{\theta}}$ estimates.

The performance of the simulated tests was examined according to two aspects: (a) accuracy and (b) item pool usage. Test accuracy was examined with the correlation between the pool raw scores and the expected scores on the tests. Pool raw scores on a domain/facet were obtained by summing the raw responses in the items of the domain/facet. Item pool usage of the tests was calculated for each facet as the percentage of items belonging to the facet that was administered to the total of simulees.

Evidence for convergent and discriminant validity of the methods. First, for each simulated 12-item test, evidence for convergent validity was obtained by computing the correlations between the expected scores on the domains and the NEO-FFI-3 raw scores. Second, as the multidimensional procedures allow to estimate the facet scores, the

intercorrelations between the expected scores on the facets were obtained for the MCAT and MCAT-B methods. For each procedure and domain, the within-domain convergent correlations between facet (expected) scores on the same domain, and the between-domain discriminant correlations between the facet (expected) scores on the domain and the facet (expected) scores on the remaining domains, were analyzed to obtain evidence for convergent and discriminant validity, respectively. Due to the convergent correlations between facets of the same domain are expected to be positive in all the cases, the average value was reported. Besides, as the discriminant correlations may take positive or negative values depending on the facets involved, the average absolute value was computed in this case. The convergent and discriminant correlations between the pool raw scores on the facets were also obtained so as to establish a baseline for comparisons.

All the statistical analyses were performed using the R (R Core Team, 2017) package *mirt* (Chalmers, 2012). The program with the CAT algorithms was developed with the package *mirtCAT* (Chalmers, 2016).

3.6 Results

3.6.1 Calibrating Each Domain Separately: Application of IRT Bifactor Model

Missing data rate for item nonresponse was very low, with a maximum value of 2%. A total of 53 out of the 360 items in the pool were excluded because they presented factor loadings below .20 on the general factor of their correspondent model. The largest number of excluded items was 18 both for Neuroticism and Conscientiousness domains. It should be mentioned that, in the case of Neuroticism, the Impulsiveness facet was excluded because it had less than 5 items after the item selection analysis. In relation to the remaining domains, 5 items were excluded in the case of Extraversion and 6 both for Openness and Agreeableness traits. The final pool was composed of 307 items, with an average number of 61 items per personality domain.

Table 3.1 shows the goodness-of-fit statistics for the bifactor solutions. Model fit for the five domains was excellent; that is, in general, all the indices had values according to the recommended criteria for good fit. Average values for the indices were: $\overline{CFI} = .96$, $\overline{TLI} = .95$, $\overline{RMSEA} = .03$, $\overline{SRMSR} = .06$.

Table 3.1. *Goodness of Fit Statistics for the Five IRT Bifactor Models in the Cross-Validation Sample (*n_2 *= 413)*

Domain	M_2^*	df	RMSEA	TLI	CFI	SRMSR
Neuroticism	1818.52	1151	.04	.95	.96	.05
Extraversion	2028.48	1406	.03	.96	.97	.06
Openness	2448.04	1688	.03	.94	.95	.06
Agreeableness	1821.12	1301	.03	.96	.97	.05
Conscientiousness	1479.74	922	.04	.95	.96	.07

Note. M_2^* = fit statistic for polytomous data of Cai and Hansen (2013); *df* = degrees of freedom of M_2^*; RMSEA: Root mean square error of approximation; TLI: Tucker-Lewis Index; CFI = Comparative Fit Index; SRMSR: Standard Root Mean Square Residual.

In the final bifactor solutions, all the item parameter estimates for the corresponding theoretical structure were significantly different from zero ($p < .05$). Table 3.2 shows the average item loadings on the general and specific factors for the five domains. The average item loadings on the general factor ranged from .43 (Agreeableness and Conscientiousness) to .51 (Extraversion). Regarding the specific factors, average item loadings ranged from .25 to .48 for Neuroticism, from .30 to .48 for Extraversion, from .18 to .62 for Openness, from .22 to .55 for Agreeableness, and from .23 to .69 for Conscientiousness. For the five bifactor solutions, the average cross-loading in absolute value was low (.04 in all the cases).

Table 3.2. *Bifactor Models for the Big Five Domains: Number of Final Items and Average Item Loadings on the General and Specific Factors*

Domain/Facets	Number of final items	Average item loadings	
		General factor	Specific factor
Neuroticism	58	.50	
Anxiety	11	.53	.25
Angry/hostility	9	.37	.48
Depression	12	.64	.33
Self-Consciousness	14	.41	.43
Vulnerability	12	.52	.43
Extraversion	64	.51	
Warmth	13	.53	.31
Gregariousness	14	.53	.30
Assertiveness	9	.54	.38
Activity	11	.47	.30
Excitement seeking	7	.41	.48
Positive emotions	10	.57	.47
Openness	69	.45	
Fantasy	13	.39	.62
Aesthetics	13	.51	.40
Feelings	9	.38	.48
Actions	13	.43	.48
Ideas	11	.56	.18
Values	10	.39	.45
Agreeableness	62	.43	
Trust	12	.29	.55
Straightforwardness	9	.39	.49
Altruism	12	.61	.25
Compliance	8	.40	.33
Modesty	10	.31	.52
Tender-Mindedness	11	.52	.22
Conscientiousness	54	.43	
Competence	8	.40	.53
Order	11	.42	.55
Dutifulness	5	.40	.38
Achievement striving	11	.45	.33
Self-discipline	11	.54	.23
Deliberation	8	.32	.69

3.6.2 Degree of Essential Unidimensionality of the Domains

The *ECV*, *PUC*, and *RB* values for the five bifactor solutions are presented in Table 3.3. The average *ECV* for the five domains was .52. This indicates that, overall, the general factor explains about 52% of the common variance, whereas approximately 48% of the common variance is distributed across the specific factors in the five domains. Extraversion showed the highest value (*ECV* = .62), whereas Conscientiousness yielded the lowest (*ECV* = .44). The average *PUC* was .83, which indicates that the great majority of the correlations theoretically reflect the general factor in the five domains. Regarding the *RB*, only Extraversion showed low parameter bias (*RB* = 7%). For Neuroticism, the *RB* was 10%, indicating non-negligible bias. Parameter bias was severe in the case of Openness (*RB* = 16%), Conscientiousness (*RB* = 16%), and Agreeableness (*RB* = 19%). It should be noted that lower *RB* values were associated with higher *ECV* values. For example, for Extraversion, which showed the highest *ECV* value, the *RB* was minor.

Table 3.3. *Explained Common Variance (ECV), Percentage of Uncontaminated Correlations (PUC), and Relative Bias (RB) for the Bifactor Models*

Domain	ECV	PUC	RB (%)
Neuroticism	.58	.81	10
Extraversion	.62	.84	7
Openness	.46	.84	16
Agreeableness	.50	.84	19
Conscientiousness	.44	.84	16

3.6.3 Precision and Evidence for Convergent Validity for Pool Scores

The alpha coefficient for the pool scores on the domains was excellent, with values that ranged from .92 (Conscientiousness) to .95 (both for Neuroticism and Extraversion). Both for Openness and Agreeableness, α was .93. Regarding the precision of the pool facet scores, almost all alpha values were above .70, except for the case of Compliance (α = .67) and Dutifulness (α = .60) facets. Values for the facets of each domain ranged from .81

(Angry/hostility) to .91 (Depression) for Neuroticism, from .79 (Excitement seeking) to .90 (Positive emotions) for Extraversion, from .76 (Feelings) to .91 (Fantasy) for Openness, from .67 (Compliance) to .86 (Altruism) for Agreeableness, and from .60 (Dutifulness) to .89 (Order, Deliberation) for Conscientiousness.

Figure 3.2 illustrates the *SE* for the IRT θ estimates in the general domains of the bifactor solutions when the complete pool is administered. For trait estimates between –3 and 3, the *SE* was lower than .40 for the five domains, which is approximately equivalent to a reliability coefficient of .84. On average, the lowest *SE*s were for Extraversion ($\overline{SE}$ = .26) and Neuroticism ($\overline{SE}$ = .27) whereas Conscientiousness showed the largest value ($\overline{SE}$ = .34). For Openness and Agreeableness, the $\overline{SE}$ was .32. This indicates that the item pool calibrated according to the bifactor model provides excellent information across the different trait levels of each domain.

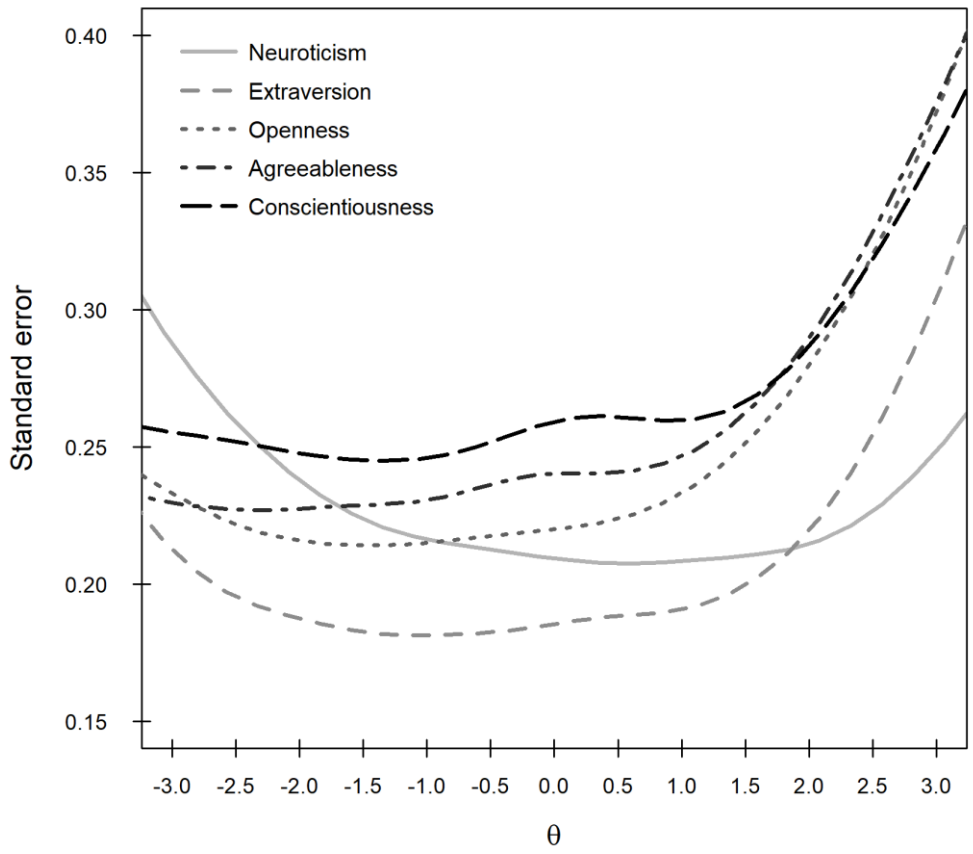

Figure 3.2. Standard error across pool domain scores for the Big Five traits.

Regarding the convergent validity between the pool scores on the domains and scores on the NEO-FFI-3 scales, the degree of association was excellent for the five traits. Neuroticism and Extraversion showed the highest values (in both cases, r = .90) whereas the lowest values were for Agreeableness and Conscientiousness (r = .83). For Openness, r = .85.

3.6.4 Post-Hoc Simulation Study

Table 3.4 shows the correlations between the pool raw scores and IRT expected scores on the domain and facets for the four simulated tests. At the domain level, the multidimensional tests (e.g., MCAT and MCAT-B) showed the best performance with the highest correlations on average ($\bar{r}$ = .94 and .93, respectively), whereas the unidimensional procedures (e.g., short scale and UCAT) were generally less accurate ($\bar{r}$ = .89 for both methods). The MCAT and MCAT-B tests performed similarly across the five domains (e.g., for Neuroticism, r = .94 for both methods), and UCAT and the short scale showed similar results (e.g., for Extraversion, r = .95 for the two tests). Taking this into account, the results for each domain are summarized by comparing the correlations of the MCAT-B and UCAT procedures. Both tests showed statistically significant differences (p<.001) in performance in favor of MCAT-B for Agreeableness, Openness, Conscientiousness, and Neuroticism ($r_{\text{MCAT-B}}$-r_{UCAT} = .09, .05, .05, and .04, respectively). Only in the case of Extraversion did both tests perform similarly. These differences in performance are consistent with previous results regarding the essential unidimensionality of the domains. Thus, Conscientiousness, Openness, and Agreeableness, which showed the lowest *ECVs* (.44, .46, and .50, respectively), also presented the highest parameter biases when a unidimensional model was fit to the data (*RB* = 16, 16, and 19%, respectively) and, therefore, the highest differences in performance between MCAT-B and UCAT. In the case of Extraversion, this domain presented the highest *ECV* (.62) and the lack of differences between UCAT and MCAT-B is, in turn, consistent with the slight bias (*RB* = 7%) found when a one-factor model was fit to the data.

Table 3.4. *Pearson Correlations between the Pool Domain/Facet Scores and Expected Scores on the Big Five Domain and Facets for the Short Scale, UCAT, MCAT, and MCAT-B*

Domain/Facet	Short scale	UCAT	MCAT	MCAT-B
Neuroticism	.89	.90	.94	.94
Anxiety			.86	.89
Angry/hostility			.86	.84
Depression			.93	.91
Self-Consciousness			.86	.85
Vulnerability			.90	.88
Extraversion	.95	.95	.96	.95
Warmth			.90	.89
Gregariousness			.86	.87
Assertiveness			.91	.89
Activity			.83	.83
Excitement seeking			.90	.90
Positive emotions			.92	.92
Openness	.89	.88	.94	.93
Fantasy			.89	.88
Aesthetics			.90	.89
Feelings			.80	.80
Actions			.85	.87
Ideas			.85	.83
Values			.84	.84
Agreeableness	.85	.83	.93	.92
Trust			.86	.85
Straightforwardness			.85	.84
Altruism			.89	.88
Compliance			.79	.78
Modesty			.84	.83
Tender-Mindedness			.84	.85
Conscientiousness	.87	.88	.93	.93
Competence			.87	.85
Order			.90	.92
Dutifulness			.78	.78
Achievement striving			.85	.85
Self-discipline			.85	.84
Deliberation			.91	.90

Note. UCAT: Unidimensional computerized adaptive test; MCAT: Multidimensional computerized adaptive test; MCAT-B: Multidimensional computerized adaptive test based on the bifactor model.

On another hand, at the facet-level, the MCAT and the MCAT-B procedures revealed a similar performance on average across the five domains: for Neuroticism $\bar{r}_{MCAT} = .88$ and $\bar{r}_{MCAT\text{-}B} = .87$, for Extraversion $\bar{r}_{MCAT} = .89$ and $\bar{r}_{MCAT\text{-}B} = .88$, for Openness $\bar{r}_{MCAT} = .86$ and $\bar{r}_{MCAT\text{-}B} = .85$, for Agreeableness $\bar{r}_{MCAT} = \bar{r}_{MCAT\text{-}B} = .84$, and for Conscientiousness $\bar{r}_{MCAT} = \bar{r}_{MCAT\text{-}B} = .86$.

Figure 3.3 shows the percentage of items belonging to each facet that was administered in each simulated test to assess the domains. In the case of the short scale, all the respondents answered the same items, which were those with the highest loadings when applying UIRT. As the items are selected according to their loading on the one-factor model, there are a different number of items for each facet, and sometimes, a facet is not even measured in the short scale. The same thing occurred with UCAT because the most informative items are selected. This explains the heterogeneous representation of the facets across the five domains for the short scales and UCAT. Indeed, the facets with the highest percentages of representation were the same when using the short scale and UCAT. On the contrary, in the case of MCAT and MCAT-B, all the facets were represented to a similar degree. For example, in the case of Extraversion, the percentage of items belonging to each facet ranged from 13 to 20% in the MCAT and from 13 to 24% in the case of the MCAT-B. This indicates that the multidimensional approaches provide a better content-balance strategy than the unidimensional ones. It should be noted that, in the case of Extraversion, which was shown to be the most unidimensional domain, the distributions for the short scale and the UCAT tended to be more uniform; that is, more similar to the distributions of the multidimensional tests than were observed in the remaining domains.

3.6.5 Evidence for Convergent and Discriminant Validity of the Methods

The results for the convergent validity with the NEO-FFI-3 scales are shown in Table 3.5. For Agreeableness, which proved to be one of the most multidimensional constructs (i.e., the one

which the highest *RB*), the multidimensional procedures showed stronger convergence (e.g., $r_{MCAT} = .79 > r_{UCAT} = .65$). In contrast, in the case of Extraversion, which was the most unidimensional domain, the four procedures showed slight differences in performance (i.e., the greater difference was .02). For the remaining domains, the differences between tests were also small (i.e., the greater difference was .05) and the evidence was mixed. For Neuroticism and Openness, the unidimensional procedures showed stronger convergence than the multidimensional procedures (e.g., for Neuroticism, $r_{UCAT} = .90 > r_{MCAT\text{-}B} = .86$), whereas for Conscientiousness all the CATs showed similar performance and better convergence than the short test (e.g., $r_{short} = .74 < r_{MCAT\text{-}B} = .79$).

The results of the analysis of the convergent and discriminant correlations between facets for the MCAT, the MCAT-B, and the item pool are shown in Table 3.6. Regarding the within-domain convergent correlations (Table 3.6, top), they were systematically higher on average for the facets of those domains that proved to be more unidimensional and lower for the facets of the domains that showed a more multidimensional structure (e.g., with the pool raw scores, the highest average correlation was .53 for Extraversion, whereas the lowest was .32 for Conscientiousness). Regarding the methods, both multidimensional tests produced an overestimation of the correlations that was slightly higher in the case of the MCAT-B.

As expected, the discriminant correlations (Table 3.6, bottom) were lower than the convergent correlations (e.g., for Extraversion, the average absolute discriminant *r* for the MCAT-B was .21 whereas the average convergent *r* was .72). This indicates that the facets of a domain were well differentiated from the facets of other domains. Both the MCAT and the MCAT-B performed similarly across the five domains.

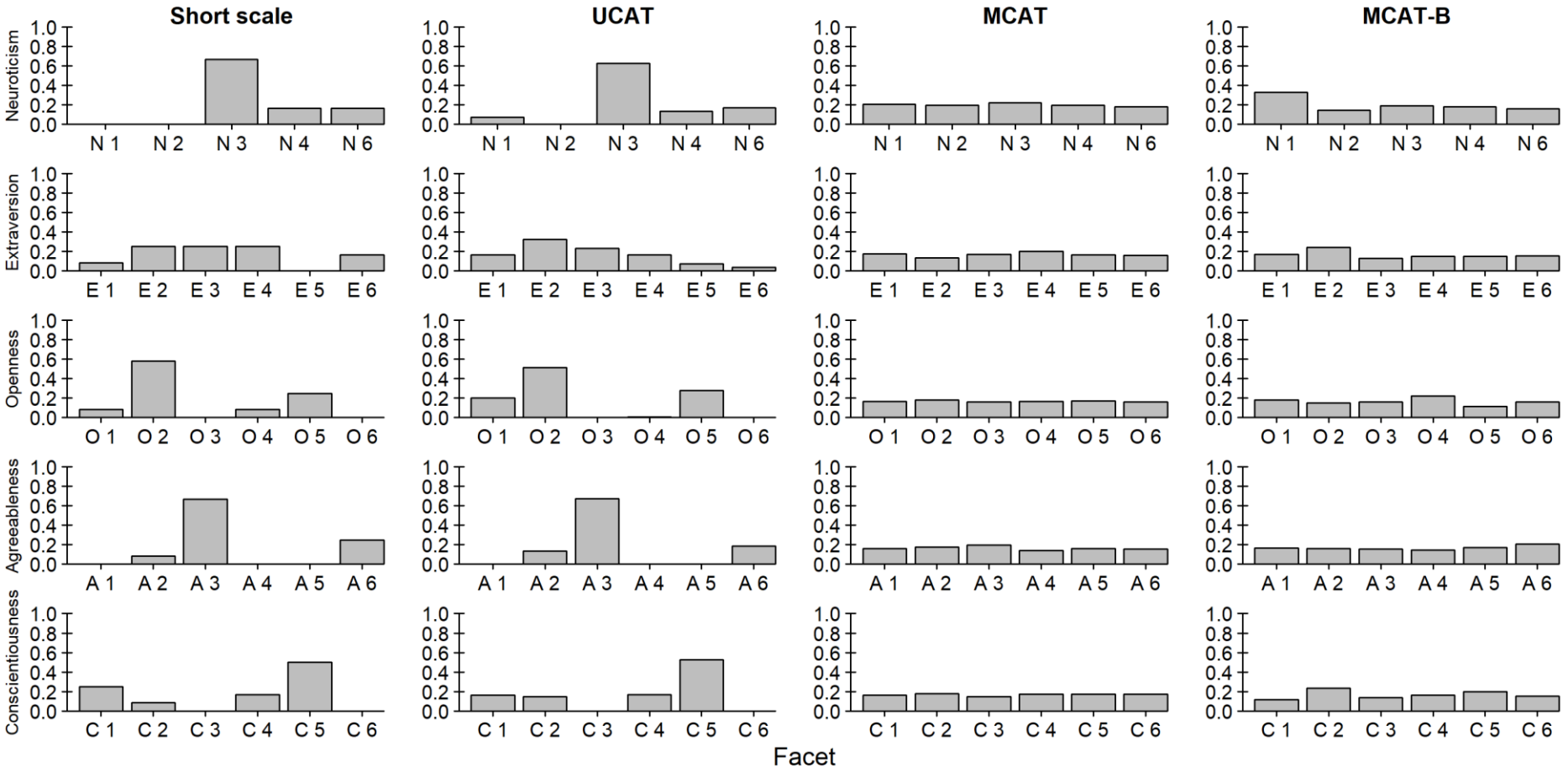

Figure 3.3. Rate of items selected from each specific facet in the four types of tests for each personality domain. UCAT: Unidimensional computerized adaptive test; MCAT: Multidimensional computerized adaptive test; MCAT-B: Multidimensional computerized adaptive test with bifactor model; N1,…, N6: Facets of Neuroticism; E1,…, E6: Facets of Extraversion; O1,…, O6: Facets of Openness; A1,…, A6: Facets of Agreeableness; C1,…, C6: Facets of Conscientiousness.

Table 3.5. *Convergent Validity with the NEO-FFI-3 for the Short Scale, UCAT, MCAT, and MCAT-B*

Test	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
Short scale	.89	.87	.83	.65	.74
UCAT	.90	.88	.85	.65	.78
MCAT	.87	.87	.82	.79	.77
MCAT-B	.86	.86	.81	.77	.79
Item pool	**.90**	**.90**	**.85**	**.83**	**.83**

Note. The values for the 307-item pool are shown in boldface. UCAT: Unidimensional computerized adaptive test; MCAT: Multidimensional computerized adaptive test; MCAT-B: Multidimensional computerized adaptive test based on the bifactor model.

Table 3.6. *Convergent and Discriminant Correlations for the Item Pool, MCAT, and MCAT-B*

Test	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
			Average within-domain convergent correlation		
MCAT	.63	.69	.49	.59	.46
MCAT-B	.67	.72	.54	.59	.52
Item pool	**.51**	**.53**	**.35**	**.41**	**.32**
			Average absolute between-domain discriminant correlation		
MCAT	.20	.22	.12	.12	.15
MCAT-B	.20	.21	.10	.12	.14
Item pool	**.20**	**.22**	**.13**	**.14**	**.17**

Note. The average within-domain convergent correlation refers to the average value of the individual correlations between the facet (expected) scores on the same domain. The average absolute between-domain discriminant correlation refers to the average value of the individual correlations (in absolute value) between the facet (expected) scores on one domain and the facet (expected) scores on the remaining domains. The values for the 307-item pool are shown in boldface. MCAT: Multidimensional computerized adaptive test; MCAT-B: Multidimensional computerized adaptive test based on the bifactor model.

3.7 Discussion

The purpose of this study was to examine whether a MCAT-B can more efficiently provide estimates of the Big Five traits than three other competing approaches: UCAT, MCAT with correlated factors, and a short scale. For the five domains, the estimated bifactor model with a general factor representing the domain and several specific factors representing the corresponding facets fit the data well. In addition, convergent validity between the calibrated pool and the NEO-FFI-3 questionnaire was excellent for the five domains. When the essential unidimensionality of the domains was tested, the *PUC* was high in all the cases, that is, the influence of the specific facets was low in the factor structure, but the *ECV* did not suggest the presence of a sufficiently strong general factor. Extraversion obtained the *ECV* value closest to .70 (*ECV* = .62), closely followed by Neuroticism (*ECV* = .58). Although both domains showed similar *ECV* values, the parameter bias was higher for Neuroticism. The remaining domains showed severe parameter bias, with *RB* values greater than or equal to 16% when UIRT was applied. Taking this into account, none of the domains clearly showed a strong unidimensional structure. However, Extraversion was the one that obtained the lowest parameter bias, so it is the only domain whose structure is closer to unidimensionality.

The results from the post-hoc simulation study revealed that, generally, for each domain, the unidimensional instruments (i.e., the short scale and UCAT) showed a similar performance, worse than did the multidimensional procedures (i.e., MCAT and MCAT-B). Specifically, results for each domain were closely related to its degree of essential unidimensionality. Thus, only in the case of Extraversion, which was the most unidimensional domain, the short scale and UCAT were shown to be as efficient as the multidimensional procedures in recovering the domain scores. Regarding the pool usage of UCAT for the five domains, there is a greater representation of the facets with a greater number of highly informative items, whereas few or no items were selected for the remaining facets. This is

consistent with the representation of the facets for the short scales, which were composed of the 12 best items in the UIRT model (i.e., the items with the highest factor loadings). These results are in line with the findings of Reise and Henson (2000), who concluded that similar results can be found using UCAT and the best items (i.e., the most informative) of a scale, although they referred to the unidimensional evaluation of the Big Five facets. It should be noted that, only for Extraversion, the distributions for the short scale and the UCAT tended to be more uniform; that is, both instruments tended to better balance the representation of facets. Despite this, for the UCAT, the content balance remained disproportionate in favor of some facets (e.g., Gregariousness) and the short scale did not contain any item from the facet of Excitement seeking. Misrepresentation of facets has been targeted as a limitation of the use of short scales because it can constitute a source of model misfit (Gignac et al., 2007). In the case of UCAT, the item pool usage could be improved by setting content constraints for the specific facets (Makransky et al., 2012). In this regard, the MCAT and MCAT-B methods showed a clear advantage in terms of balancing pool usage not only for Extraversion but for all the Big Five domains, so that items from all the facets were always administered in similar proportions.

For the domains which proved to be more multidimensional (Neuroticism, Agreeableness, Openness, and Conscientiousness), the MCAT and MCAT-B methods outperformed the unidimensional procedures when estimating the domain scores. Regarding the recovery of the pool facet scores, both procedures showed a similar performance across the five domains. Besides, according to the evidence of validity obtained in this study, at the domain level, the multidimensional methods presented greater validity for the Agreeableness domain, which showed to be the most multidimensional according to the *RB*. For the remaining domains, the differences between procedures were generally small and in some cases (i.e., for Neuroticism and Openness) favored the unidimensional methods. This unexpected advantage might be due to the fact that the criteria (i.e., the NEO-FFI scores) are brief measures that are

directly designed to measure the domain factors, as is the case of the short test and the UCAT. In contrast, the goal in a multidimensional CAT is to recover not only the domain scores but also the facet scores. In this sense, whereas the better content balance of multidimensional tests led to a better recovering of the pool raw score, it might also be slightly reducing the efficiency for measuring the general domain. Consistent with this, the differences between procedures were smaller for the Extraversion domain, in which the unidimensional tests achieved a good content balance.

At the facet level, both multidimensional methods performed similarly and only showed slight differences in the convergent and discriminant correlations. It must be noted that some inflation was found for the within-domain convergent correlations (i.e., between facets of the same domain). This overestimation might be partly due to the bias in the Bayesian estimates produced by the inclusion of the prior correlation matrix, and thus caution should be exercised when interpreting these correlations between estimates (Segall, 1996).

Despite the similarities between the multidimensional tests, the use of bifactor modeling offers several advantages over the correlated-factor model, which make it a more desirable approach to assess multifaceted constructs. It should be noted that this advantages are not inherent to CAT. First, as we have illustrated in this study, some bifactor-derived indices (i.e., *ECV* and *PUC*) can be easily obtained by researchers to examine the degree of unidimensionality of the constructs in order to determine whether a UIRT or MIRT model is required (e.g., Rodríguez et al., 2016a, 2016b). Second, the bifactor model yields an estimated score in the general domain with an associated standard error (*SE*), which is an indicator of the accuracy of the overall measure. Although in MIRT with correlated factors, a general score in the domain can be obtained by averaging the results over the specific facets (e.g., Makransky et al., 2012), this cannot be directly estimated and, therefore, the model does not provide any information on its accuracy. A third advantage not explored here is that it also allows estimating

the accuracy associated with the IRT facet scores. Although this residualized facet scores are difficult to interpret, many studies have shown how they can contribute to the incremental prediction of several psychological measures above and beyond scores on the general factor. For example, McAbee et al. (2014) applied separate bifactor models for each of the six traits of the HEXACO model of personality (Lee & Ashton, 2004) and examined the role of the general factor and the specific facet scores for predicting students' performance. They concluded that modeling facet scores enables researchers to explain interesting but complex relations between narrow personality traits and student performance outcomes, which cannot be otherwise studied. In addition, facet scores may be especially informative for assessment contexts where individual personality profiles need to be developed attending not only to the broader trait (i.e., the domain) but also to the individual differences reflected by the narrower traits (i.e., the facets). Related to this, it is important to note that several authors have highlighted the importance of evaluate whether subscale scores in multidimensional measurement models have added value over the total score and, therefore, if they should be computed and used (e.g., Reise et al., 2013; Sinharay, Puhan, & Haberman, 2011; Sinharay, 2013). In the case of the bifactor model, some indices that can be applied to subscales, such as omega and omega hierarchical, have been used for this purpose (Rodríguez et al., 2016a, 2016b).

Besides, a number of studies have reported great gains in efficiency associated with the use of MCAT-B when estimating domain and facets scores (e.g., Gibbons et al., 2012). In this study, the time required to complete the final 307-item pool was 62.23 minutes approximately and, proportionally, about 12.16 minutes to complete any of the 60-item adaptive versions. This supposes important reductions of testing time and test length (i.e., 83%). Moreover, considering that in this short time the MCAT-B procedure provides both the domain and facet scores, the advantage over UCAT (and the short scale), which only provides the domain score,

is evident when facet scores are required, for example, for diagnosis purposes. Likewise, as mentioned above, in such time the use of the bifactor model allows to obtain a measure of precision of the domain score (i.e., the *SE*) which cannot be obtained with the MCAT procedure. This is especially relevant when evaluating multidimensional constructs and the objective of the evaluation is to provide the domain score. For example, Moore et al. (in press) applied the bifactor model to design a CAT to measure the general trait of schizotypal personality, which includes several features or dimensions (e.g., cognitive-perceptual). As these authors pointed out, in these cases the bifactor model allows to account for multidimensionality through the inclusion of the specific factors, which in fact contribute to the measurement precision of the general domain. Moreover, fitting the unidimensional IRT model to multidimensional data would not be optimal either because it may lead to biased item parameter estimates (Reise, Moore, & Maydeu-Olivares, 2011; Reise, Cook, & Moore, 2015).

Previous studies assessing the Big Five model with MCAT (Makransky et al., 2012) or applying MCAT-B (Gibbons et al., 2008, 2012, 2014, 2016; Moore et al., in press; Sunderland et al., 2017; Weiss & Gibbons, 2007; Zheng et al., 2013) specified confirmatory structures to calibrate the item pools. In the current study, we illustrated the application of more realistic bifactor exploratory models to measure the Big Five adaptively.

The current study has several limitations that deserve further discussion. First, we are aware of the problems of the generalizability of the findings to other contexts due to the specificity of the study sample. In this regard, examining the intercorrelations between the five personality factors, we have found they are consistent with previous research. For example, Neuroticism correlated negatively with the remaining domains and showed high associations with Extraversion (r = -.56) and Conscientiousness (r = -.19), whereas domains such as Openness and Agreeableness showed lower correlations (r = .13; Mount, Barrick, Scullen, & Rounds, 2005; Van der Linden, te Nijenhuis, & Bakker, 2010). Therefore, although in this

study, the pattern of relationships between the Big Five domains is similar to that previously found, further research is required to replicate these results in other sub-populations. Second, a post-hoc simulation was conducted to examine the performance of the tests, and therefore simulees' responses were drawn from the real dataset. Although real data simulations are essential to evaluate how CAT procedures will operate with real respondents (Thompson & Weiss, 2011), it is necessary to carry out additional studies with live examinees to investigate their performance in real testing settings. Third, we have defined the adaptive algorithms according to a unique item selection criterion (i.e., D-Optimality). Future research should evaluate the performance of alternative item administration criteria. For example, Seo and Weiss (2015) showed through a Monte Carlo simulation study that the D_s-Optimality criterion worked well when the focus is on measuring the general factor of a bifactor model, whereas other rules such as D- or A-optimality improved the measurement of the specific factors.

In closing, this study provides two main contributions to previous research concerning the adaptive assessment of personality. First, the Big Five domains are essentially multidimensional constructs and, therefore, they cannot be adequately evaluated through the application of a unidimensional model. Second, and related to the previous conclusion, MCAT-B constitutes a preferential framework for adaptively assessing the Big Five of personality because it allows assessing the general domains while representing the multidimensionality due to the specificity of the facets. Several other applications of the bifactor model have been illustrated to address a number of issues of interest in personality research. In this regard, it is common to observe how individual differences in the response style constitute a source of variance that can systematically distort the factor structure of personality instruments and lead to model misfit (e.g., Podsakoff, MacKenzie, & Podsakoff, 2012). Abad et al. (2016) illustrated how the inclusion of an acquiescence method factor can be a useful tool to separate variance explained by general and specific traits of personality from variance due to the acquiescent

response style. It would also be interesting to include social desirability item markers to study its relationship with the Big Five domains and facets and to determine how this response style can affect the prediction of different psychological constructs (see, e.g., Ferrando, Lorenzo-Seva, & Chico, 2009). Taking all the above into account, future research in the area of adaptive assessment of personality should be oriented toward the modeling and study of response styles during the phases of calibration and administration of the item pool.

References

Abad, F. J., Sorrel, M. A., García, L. F., & Aluja, A. (2016). Modeling general, specific, and method variance in personality measures: Results for ZKA-PQ and NEO-PI-R. *Assessment*, 1-19. doi:10.1177/1073191116667547

Ashton, M. C., Paunonen, S. V., & Lee, K. (2014). On the validity of narrow and broad personality traits: A response to Salgado, Moscoso, and Berges (2013). *Personality and Individual Differences, 56*, 24-28. doi:10.1016/j.paid.2013.08.019

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443-459. doi:10.1007/BF02293801

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling, 22,* 504-516. doi:10.1080/10705511.2014.938596

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hasings Robbins-Monro algorithm. *Psychometrika, 75*, 33-57. doi: 10.1007/s11336-009-9136-x

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*, 307-335. doi: 10.3102/1076998609353115

Cai, L., & Hansen, M. (2013). Limited- information goodness- of- fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66,* 245-276. doi:10.1111/j.2044-8317.2012.02050.x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1-29. doi:10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*, 1-39. doi:10.18637/jss.v071.i05

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality, 80,* 219-251. doi:10.1111/j.1467-6494.2011.00739.x

Cordero, A., Pamos, A., & Seisdedos, N. (2008). *Revised NEO Personality Inventory (NEO PI-R) manual. Spanish adaptation. Madrid: TEA Ediciones.*

Costa, P., & McCrae, R. R. (1992). *NEO PI-R manual professional.* Odessa, FL: Psychological Assessment Resources, Inc.

DeMars, C. E. (2013) A Tutorial on Interpreting bifactor model scores. *International Journal of Testing, 13,* 354-378, DOI: 10.1080/15305058.2013.799067

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2009). A general factor-analytic procedure for assessing response bias in questionnaire measures. *Structural Equation Modeling, 16,* 364-381. doi:10.1080/10705510902751374

Forbey, J. D., & Ben-Porath, Y. S. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 computerized adaptive version. *Psychological Assessment, 19,* 14-24. doi: 10.1037/1040-3590.19.1.14

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31,* 4-19. doi: 0.1177/0146621606289485

Gibbons, R. D., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2014). Development of the CAT-ANX: A computerized adaptive test for anxiety. *American Journal of Psychiatry, 171*, 187-194. doi: 10.1176/appi.ajp.2013.13020178

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology, 12*, 83-104. doi: 10.1146/annurev-clinpsy-021815-093634

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*, 361-368. doi:10.1176/appi.ps.59.4.361.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *Archives of General Psychiatry*, *69*, 1104-1112. doi:10.1001/archgenpsychiatry.2012.14

Gignac, G. E., Bates, T. C., & Jang, K. L. (2007). Implications relevant to CFA model misfit, reliability, and the five-factor model as measured by the NEO-FFI. *Personality and Individual Differences*. *43*, 1051-1062. doi:10.1016/j.paid.2007.02.024

Goldberg, L. R. (1999). A broad-band width, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe*, *7*, 7-28. Tilburg, The Netherlands: Tilburg University Press.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi:10.1080/10705519909540118

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. John, R. Robins, &

L. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). New York, Guilford.

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51*, 78-89. doi:10.1016/j.jrp.2014.05.003

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate behavioral research, 39*, 329-358. doi: 10.1207/s15327906mbr3902_8

Makransky, G., Mortensen, E. L., & Glas, C. A. (2012). Improving personality facet scores with multidimensional computer adaptive testing: An illustration with the NEO PI-R. *Assessment, 20*, 3-13. doi:10.1177/1073191112437756

Maples, J. L., Guan, L., Carter, N. T., & Miller, J. D. (2014). A test of the International Personality Item Pool representation of the Revised NEO Personality Inventory and development of a 120-item IPIP-based measure of the five-factor model. *Psychological Assessment, 26*, 1070-1084. doi: 10.1037/pas0000004

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of IRT and factor analysis models. *Structural Equation Modeling, 18*, 333–356. doi:10.1080/10705511.2011.581993

McAbee, S. T., Oswald, F. L., & Connelly, B. S. (2014). Bifactor models of personality and college student performance: A broad versus narrow view. *European Journal of Personality, 28*, 604-619. doi:10.1002/per.1975

McCrae, R. R., & Costa Jr., P. T. (2007). Brief versions of the NEO-PI-3. *Journal of Individual Differences, 28*, 116-128. doi:10.1027/1614-0001.28.3.116

McCrae, R. R., Costa, Jr, P. T., & Martin, T. A. (2005). The NEO–PI–3: A more readable revised NEO personality inventory. *Journal of Personality Assessment, 84*, 261-270. doi: 10.1207/s15327752jpa8403_05

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64-82. doi:10.1037/1082-989X.7.1.64

Moore, T. M., Calkins, M. E., Reise, S. P., Gur, R. C., & Gur, R. E. (in press). Development and Public Release of a computerized adaptive (CAT) version of the Schizotypal Personality Questionnaire. *Psychiatry Research*. doi: 10.1016/j.psychres.2018.02.022

Mount, M. K., Barrick, M. R., Scullen, S. M., & Rounds, J. (2005). Higher-order dimensions of the big five personality traits and the big six vocational interest types. *Personnel Psychology*, *58*, 447-478. doi:10.1111/j.1744-6570.2005.00468.x

Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52,* 431–462. doi: 10.1007/BF02294365

Nieto, M. D., Abad, F. J., Hernández-Camacho, A., Garrido, L. E., Barrada, J. R., Aguado, D., & Olea, J. (2017). Calibrating a new item pool to adaptively assess the Big Five. *Psicothema*, *29*, 390-395. doi:10.7334/psicothema2016.391

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual differences*, *43*, 971-990. doi:10.1016/j.paid.2007.03.017

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*, 539-569. doi: 10.1146/annurev-psych-120710-100452

Quinn, H. O. (2014). *Bifactor models, explained common variance (ECV), and the usefulness of scores from unidimensional item response theory analyses*. Unpublished Master's thesis, The University of North Carolina at Chapel Hill, NC.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667-696. doi:10.1080/00273171.2012.715555

Reise, S.P., Cook, K.F., Moore, T.M. (2015). Evaluating the impact of multidimensionality on unidimensional Item Response Theory model parameters, in: Reise, S.P., Revicki, D. (Eds.), Handbook of Item Response Theory Modeling. Routledge, New York, NY.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*, 347-364. doi:10.1177/107319110000700404

Reise, S. P., Moore, T., Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement*. *71*, 684-711. doi: 0.1177/0013164410378690

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73*, 5-26. doi: 10.1177/0013164412449831

Rodríguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*, 223-237. doi:10.1080/00223891.2015.1089249

Rodríguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*, 137-150. doi: 10.1037/met0000045

Rudick, M. M., Yam, W. H., & Simms, L. J. (2013). Comparing countdown-and IRT-based approaches to computerized adaptive personality testing. *Psychological Assessment, 25*, 769-779. doi: 10.1037/a0032541

Salgado, J. F., Moscoso, S., Sanchez, J. I., Alonso, P., Choragwicka, B., & Berges, A. (2014). Validity of the five-factor model and their facets: The impact of performance measure and facet residualization on the bandwidth-fidelity dilemma. *European Journal of Work and Organizational Psychology, 24*, 325- 349. doi:10.1080/1359432X.2014.903241

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354. doi: 10.1007/BF02294343

Seo, D. G., & Weiss, D. J. (2015). Best design for multidimensional computerized adaptive testing with the bifactor model. *Educational and Psychological Measurement, 75*, 954-978. doi: 10.1177/0013164415575147

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment, 17*, 28–43. doi: 10.1037/1040-3590.17.1.28

Simms, L. J., Prisciandaro, J. J., Krueger, R. F., & Goldberg, D. P. (2012). The structure of depression, anxiety and somatic symptoms in primary care. *Psychological Medicine, 42*, 15-28. doi: 10.1017/S0033291711000985

Sinharay, S. (2013). A note on assessing the added value of subscores. *Educational Measurement: Issues and Practice, 32*, 38-42. doi:10.1111/emip.12021

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30,* 29-40. doi:10.1111/j.1745-3992.2011.00208.x

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107-120. doi: 10.1007/S11336-008-9101-0

Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp.183_206). New York, NY: Routledge/Taylor & Francis Group.

Sunderland, M., Batterham, P., Carragher, N., Calear, A., & Slade, T. (2017). Developing and validating a computerized adaptive test to measure broad and specific factors of internalizing in a community sample. *Assessment,* 1-16. doi: 1073191117707817.

Ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69,* 613-625. doi: 10.1007/BF02289858

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16,* 1-9. Retrieved from http://pareonline.net/getvn.asp?v=16&n=1.

Van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44,* 315-327. doi:10.1016/j.jrp.2010.03.003

Weiss, D. J., & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC conference on computerized adaptive testing*.

Zheng, Y., Chang, C. H., & Chang, H. H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research, 22,* 491-499. doi:10.1007/s11136-012-0179-6

Chapter 4

Does Modeling Wording Effects Help Recover Uncontaminated Person Scores? A Systematic Evaluation with Random Intercept Item Factor Analysis

Abstract

The item wording (or keying) effect consists of logically inconsistent answers to positively and negatively worded items that tap into similar (but polarly opposite) content. Previous research has shown that this effect can be successfully modeled through the random intercept item factor analysis (RIIFA) model, as evidenced by the improvements in model fit in comparison to models that only contain substantive factors. However, little is known regarding the capability of this model in recovering the uncontaminated person scores. To address this issue, the current study analyzed for the first time the performance of the RIIFA approach across three types of wording effects proposed in the literature: carelessness, item verification difficulty, and acquiescence. In the context of unidimensional substantive models, four independent variables were manipulated using Monte Carlo methods: type of wording effect, amount of wording effect, sample size, and test length. The results corroborated previous findings by showing that the RIIFA models were consistently able to account for the variance in the data, attaining excellent fit regardless of the amount of bias. Conversely, the models without the RIIFA factor produced increasingly poorer fit with greater amounts of wording effects. Surprisingly, however, the RIIFA models were not able to better estimate the uncontaminated person scores for any type of wording effect in comparison to the substantive unidimensional models. These apparently paradoxical findings are explained in light of the properties of the factor models examined.

4.1 Introduction

Most self-report scales in Psychology often include both positively worded (PW) items, which are intended to measure the presence of a construct with positive valence (e.g., Extraversion), and negatively worded (NW) items, which measure the presence of a construct with negative valence (e.g., Introversion; Kam & Meyer, 2015a; Kam, 2016, 2018). The goal of this practice is usually to measure the two poles of the same construct. For example, a scale measuring Extraversion may include several PW items (e.g., I make friends easily) as well as some NW items measuring Introversion (e.g., I prefer to be alone), which taps the polar opposite end of the construct. However, when both types of items are combined, respondents may manifest differential response styles to PW and NW items. This phenomenon is known as the *item wording effect* and consists of logically inconsistent answers to PW and NW items that tap into similar (but polar opposite) content (Kam & Meyer, 2015a, Kam, 2016).

For decades, the presence of different wording effects has been ubiquitous in psychological measurement (Carmines & Zeller, 1979; Johnson & Messick, 1958; Paulhus, 1991). An extensive body of research has demonstrated that wording effects may impact the psychometric properties of scales, deteriorating model fit (Abad, Sorrel, García, & Aluja, 2018; Danner, Aichholzer, & Rammstedt, 2015; Woods, 2006), spuriously increasing the dimensionality due to the emergence of separate factors for PW and NW items (Barnette, 2000; Marsh, 1996; Rodebaugh et al., 2004; Schmitt & Stults, 1985), reducing the reliability of measures (Roszkowski & Soven, 2010; Schriesheim, Eisenbach, & Hill, 1991), inflating or suppressing the structural relationships (Kam & Meyer, 2015b; Kam, Zhou, Zhang, & Ho, 2012), and distorting the factor loading structures (Navarro-González, Lorenzo-Seva, & Vigil-Colet, 2016; Savalei & Falk, 2014; Zhang, Noor, & Savalei, 2016).

However, it is striking that the influence of wording effects on person score estimates has received less attention. A possible reason is that most studies investigating wording effects

are conducted using real data collected in applied settings, making it impossible to know the uncontaminated true score of the respondents. In addition, prior simulation studies evaluating the recovery of person scores in the presence of response biases (e.g., Plieninger, 2016; Wetzel, Böhnke, & Rose, 2016; Falk & Cai, 2016) have been mainly focused on the influence of response styles such as extreme responding (i.e., tendency to select extreme response options). Related to this, in general few studies have systematically evaluated wording effects and these are very limited because they often include a single wording effect exclusively (Schmitt & Stults, 1985; Woods, 2006). Thus, the current literature lacks a systematic evaluation of the impact that different wording effects may have, as well as of the conditions under which they are most harmful.

The random intercept item factor analysis (RIIFA) model (Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006) has shown to be a promising approach for modeling method variance due to wording effects over competing approaches (Savalei & Falk, 2014). First, it is very easy to implement in practice. Second, it generally produces substantial improvements in model fit at the cost of only one degree of freedom in comparison to the "do nothing" approach (i.e., fitting a model with only substantive factors, ignoring the presence of wording effects; Abad et al., 2018; Billiet & McClendon; Kam et al., 2012; Maydeu-Olivares & Coffman, 2006; Yang et al., 2018). Third, it is robust in recovering the substantive factor loadings, even when its main assumption (i.e., equal method factor loadings across all items) is violated (Savalei & Falk, 2014). However, despite these positive characteristics there is still limited knowledge regarding its performance in estimating certain parameters such as the uncontaminated person scores in the presence of wording effects.

In the light of the aforementioned issues, the motivating goal of this study was to examine the impact of wording effects on parameter estimation, specifically person scores, in unidimensional data sets with categorical variables. To do so, we focused on three wording

effects proposed in literature: carelessness, item verification difficulty, and acquiescence (Swain, Weathers, & Niedrich, 2008). Thus, the main aim of this study was to assess the performance of the RIIFA model in estimating person scores and other substantive parameters in the presence of wording effects, and to compare it with the "do nothing approach". The rest of this section will be devoted to provide: (a) a conceptualization of the types wording effects considered in this study and the cognitive processes underlying them, (b) some examples of response patterns of the targeted wording effects, and (c) a description of the RIIFA model.

4.2 Types of Wording Effects

A response bias is any systematic tendency to answer items irrespective of their content (Paulhus, 1991, 2002). Previous literature has usually distinguished between two types of response biases: response styles and response sets (Jackson & Messick, 1958). *Response styles* refers to a systematic tendency to use or avoid some specific response categories (e.g., extreme response style or the preference for extreme categories; e.g., Wetzel et al., 2016). A number of studies have focused on demonstrating the stability of individual response styles across time and different constructs (e.g., Danner et al., 2015; Weijters, Geuens, & Schillewaert, 2010a, 2010b). In this regard, response styles have been traditionally conceptualized as response biases that are consistent across time and situations. In contrast, *response sets* have been defined as response biases that temporarily manifest in specific situations or settings (e.g., the tendency to provide a positive self-image in a personnel selection process; Jackson & Messick, 1958; Nunnally & Berstein, 1994). Wetzel, Böhnke, and Brown (2016), Van Vaerenbergh and Thomas (2013), and Ziegler (2015) provide further review of these response biases.

Within this conceptual framework, wording effects are another type of response bias which consists of logically inconsistent answers to PW and NW items that tap into similar (but polarly opposite) content (Kam & Meyer, 2015a; Kam, 2016). This study focuses on three types of wording effects proposed in the literature: carelessness, item verification difficulty,

and acquiescence. Building on the response process model developed in the survey research literature (Tourangeau, Rips, & Rasinski 2000), Swain et al. (2008) and Weijters and Baumgartner (2012) described these wording effects in terms of the cognitive processes underlying an item response. This model consists of four major steps: (a) *comprehension* (attending to the item and interpreting it), (b) *retrieval* (retrieving a relevant belief previously formed from long-term memory or transferring to working memory the information used to construct a new belief), (c) *judgement* (integrating the information retrieved previously and comparing it to the item representation), and (d) *response* (representing the answer onto the given scale and producing a response).

4.2.1 Carelessness

Several terms have been used to refer to a pattern of responding in which respondents pay insufficient attention to the items' content, such as random responding (Meade & Craig, 2012), noncontingent responding (Baumgartner & Steenkamp, 2001), inattentiveness (Johnson, 2005), or insufficient effort responding (Huang, Curran, Keeney, Poposki & DeShon, 2012). The concept of carelessness has been broadly used to refer to different random or nonrandom response patterns such as fully or partially random responding, using the same response category (i.e., straight-line responding) or response sequence, or skipping items (e.g., Swain et al., 2008; Johnson, 2005; Meade & Craig, 2012).

The current research focuses on a systematic (non-random) type of carelessness in which a respondent may answer according to the expectations that he or she has formed about what is being measured according to the questionnaire instructions or the content of the initial items (Schmitt & Stults, 1985; Weijters, Baumgartner, & Schillewaert, 2013; Woods, 2006). This type of carelessness occurs at the initial step of the response process model (Tourangeau et al., 2000), during the comprehension phase (Swain et al., 2008; Weijters & Baumgartner, 2012; Weijters et al., 2013). Authors suggesting this variant of carelessness usually associate

it to misresponses to the NW items. This is often built on the assumption that respondents may generate the expectation that items are stated affirmatively based on everyday experiences with language, and on the results from prior studies showing that most Likert type items are affirmations (Swain et al., 2008). However, we argue that these reasons do not necessarily imply that misresponses due to carelessness will only occur to the NW items. For example, if the questionnaire instructions explicitly reveal that a construct with negative valence is being measured (e.g., burnout), a careless respondent may assume that items will be NW and he or she might fail in responding to the PW items.

Previous research investigating carelessness has mainly focused on the detection of careless respondents through the use of different methods such as instructed response items, indices based on repeated responses (e.g., long-string), and factor mixture modeling (e.g., Kam & Fan, 2018; Kam & Meyer, 2015a; Meade & Craig, 2012). Two simulation studies have examined the impact of systematic carelessness (to NW items) on the factor structure of unidimensional scales. First, Schmitt and Stults (1985) found that only 10% of careless respondents was necessary for the emergence of a spurious second dimension (these authors used principal component analysis). Second, Woods (2006) reached similar conclusions in the context of confirmatory factor analysis: with only 10% of carelessness respondents, a two-factor model presented better model fit and thus was preferred over the unidimensional solution.

4.2.2 Item Verification Difficulty

Swain et al. (2008) conceptualized item verification difficulty as a type of inconsistent responding that occurs when a respondent belief about the construct being measured (i.e., his or her true trait level) mismatches the item content during the judgement phase of the response process model (Tourangeau et al., 2000). Swain et al. (2008) suggested that the item verification process can be explained according to the constituent-comparison model

(Carpenter & Just, 1975). This model postulates that a respondent difficulty to verify an item, and thus the probability of misresponding it, will depend on the complexity in comparing his/her own belief or true trait level on the construct being measured to the item content. This difficulty will depend on whether the item content is on the same pole (i.e., is truth) or on the opposite pole (i.e., is false) relative to the respondent's belief (i.e., true trait level), and whether it is affirmed or negated. According to the constituent-comparison model, a person who believes that he or she is extroverted (i.e., has a high trait level in Extraversion) will have increasing difficulty in responding the following items: "I am extroverted" (true affirmation), "I am introverted" (false affirmation), "I am not extroverted" (false negation), and "I am not introverted" (true negation). Contrarily, a person who believes that he or she is introverted (i.e., has a low trait level in Extraversion) will have increasing difficulty in responding the following items: "I am introverted" (true affirmation), "I am extroverted" (false affirmation), "I am not introverted" (false negation), and "I am not extroverted" (true negation). This model implies that a respondent will have to perform more cognitive operations to compare an item with his or her belief as the difficulty of such comparison increases. In this study, we focused on true affirmed items and false affirmed items because prior research has generally discouraged the use of negations (e.g., Weijters & Baumgartner, 2012).

Previous studies have suggested that wording effects may be related to reading ability. The studies of Marsh (1986, 1996) showed how method effects (in this case associated to NW items) were weaker for more verbally able students. Besides, Swain et al. (2008) confirmed through a series of experiments the item verification predictions made by the constituent-comparison model: inconsistent responding and difficulty to process statements linearly increased with true affirmations, false affirmations, false negations, and true negations.

4.2.3 Acquiescence

Acquiescence is the tendency to respond to items using agree categories (i.e., the positive side of the scale) irrespective of their content (e.g., Paulhus, 1991; Weijters et al. 2013; Wetzel et al., 2016). This wording effect influences the response phase (Swain et al., 2008; Weijters & Baumgartner, 2012; Weijters et al., 2013), which is the final step of the response process model (Tourangeau et al., 2000). Knowles and Condon (1999) suggest that the cognitive process underlying acquiescence can be explained according to the dual-process model of understanding (Gilbert, 1991). This model posits that initially, a respondent automatically accepts the item content (comprehension phase), and subsequently he or she can reevaluate it in order to decide whether reject it or continue accepting it (reconsideration phase). This second step implies an effort for the participant, so it can be omitted depending on his or her ability and motivation. If this occurs, a respondent will automatically agree to all items, irrespective if they are PW or NW, manifesting an acquiescent response pattern (Swain et al., 2008; Weijters & Baumgartner, 2012).

Previous studies examining the effects of acquiescence mostly do it from an empirical perspective through the computation of different measures based on the endorsement of polar opposite items (e.g., Hinz, Michalski, Schwarz, & Herzberg, 2007; Rammstedt & Farmer, 2013) or many items with heterogeneous content (Baumgartner & Steenkamp, 2001; Kam & Zhou, 2015; Weijters et al., 2013). However, some of these measures (i.e., those not based on heterogeneous items) may also reflect other wording effects such as carelessness (Weijters et al., 2013; Kam & Meyer, 2015a), leading to erroneous conclusions about the influence of acquiescence. In contrast, very few studies have examined the impact of acquiescence from the perspective of Monte Carlo simulation. Grønhaug and Heide (1992) simulated acquiescent responses to Likert type items and found that inconsistent responses might distort results from regression and factor analysis. More recently, Plieninger (2016) found that the impact of

acquiescence on reliability, validity, and scale scores estimates was greater in unbalanced scales with fewer NW items.

4.3 Illustration of Wording Effects Response Patterns

Table 4.1 present some examples of response patterns that examinees with low (top section) or high trait levels (bottom section) may show when responding to a scale with 10 items (5 PW marked as "+", and 5 NW marked as "-"). In each case, both non-reversed (left section) and reversed responses (right section) are presented. The first row always represents the uncontaminated true pattern. All response patterns correspond to an hypothetical examinee that misresponds to 50% of the items according to different wording effects: carelessness to NW items (responses to NW items were reversed), item verification difficulty (responses to PW/NW items were reversed for a person with a true low/high trait level), acquiescence (responses to PW/NW items were replaced by categories implying agreement for a person with a true low/high trait level), and disacquiescence (responses to NW/PW items were replaced by categories implying disagreement for a person with a true low/high trait level). Looking at the total raw scores (computed with reversed item responses), it can be seen that, in general, total scores for respondents with low (high) true trait levels will be upwardly (downwardly) biased in the presence of any wording effect.

It should be noted that different wording effects might produce indistinguishable observable response patterns in practice (Kam & Meyer, 2015a; Weijters et al., 2013). For example, looking at the non-reversed responses in Table 4.1, two persons with a high true trait level may present a similar response pattern if one of them responds as if all items are positively worded (carelessness to NW items) and the other one has problems to process NW items (item verification difficulty). Wording effects might also be confounded under other circumstances not illustrated in Table 4.1. For example, some types of acquiescent respondents that systematically use the highest agree category might resemble some types of careless

Table 4.1. *Examples of Response Patterns for the Wording effects of Carelessness, Item Verification Difficulty, and Acquiescence*

	Low trait level																				
Pattern	Non-reversed responses										Reversed responses										Sum
True	1	1	1	2	2	3	3	4	4	4	1	1	1	2	2	2	2	1	1	1	14
Carelessness to NW items	1	1	1	2	2	2	2	1	1	1	1	1	1	2	2	3	3	4	4	4	**25**
Item verification difficulty	4	4	4	3	3	3	3	4	4	4	4	4	4	3	3	2	2	1	1	1	**25**
Acquiescence	3	3	3	3	4	3	3	4	4	4	3	3	3	3	4	2	2	1	1	1	**23**
	High trait level																				
	Non-reversed responses										Reversed-responses										
Pattern	+	+	+	+	+	-	-	-	-	-	+	+	+	+	+	-	-	-	-	-	Sum
True	4	4	4	3	3	2	2	1	1	1	4	4	4	3	3	3	3	4	4	4	36
Carelessness to NW items	4	4	4	3	3	3	3	4	4	4	4	4	4	3	3	2	2	1	1	1	*25*
Item verification difficulty	4	4	4	3	3	3	3	4	4	4	4	4	4	3	3	2	2	1	1	1	*25*
Acquiescence	4	4	4	3	3	4	3	3	3	3	4	4	4	3	3	1	2	2	2	2	*27*

Note. Inconsistent answers appear highlighted in grey. Overestimated person scores are shown in boldface, whereas underestimated person scores are shown in underlined italics.

respondents (e.g., one displaying a straight-line responding pattern), and vice versa. However, there is an important difference in the response process: careless respondents overlook item content (the problem arises at the initial comprehension phase) whereas acquiescent ones pay attention to it (the problem occurs at the final response phase; Kam & Meyer, 2015a; Weijters et al., 2013).

4.4 The RIIFA Model

Maydeu-Olivares and Coffman (2006) introduced the RIIFA model as an extension of the common factor model that allows for the explicit modeling of consistent individual differences in the use of the response scale. In the common factor model, the response of participant *j* to item *i* (y_{ij}) can be written as:

$$y_{ij} = \mu_i + \boldsymbol{\lambda}_i' \boldsymbol{f}_j + e_{ij}, \tag{4.1}$$

where μ_i is the intercept for item *i*, $\boldsymbol{\lambda}_i$ is the vector of factor loadings for item *i*, $\boldsymbol{f}_j$ is the vector of substantive factor scores for participant *j*, and e_{ij} is the error term for participant *j* on item *i*. Assuming that the mean of the common factors and the error terms is zero, and that the error terms are uncorrelated with each other and with the common factors, the covariance matrix implied by this model ($\boldsymbol{\Sigma}_y$) is expressed as:

$$\sum\nolimits_y = \boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta} \tag{4.2}$$

where *lambda* ($\boldsymbol{\Lambda}$) is a $k \times m$ matrix of factor loadings for k variables and m common factors, *psi* ($\boldsymbol{\Psi}$) is a $m \times m$ covariance matrix of the common factors, and *theta* ($\boldsymbol{\Theta}$) is a $k \times k$ covariance matrix of the error terms.

In the RIIFA model the intercept (γ_{ij}) is decomposed into a fixed part (μ_i) common to all respondents but differing across items, and a random part (ζ_j) common to all items but differing across respondents:

$$y_{ij} = \gamma_{ij} + \boldsymbol{\lambda}_i' \boldsymbol{f}_j + e_{ij} \qquad \gamma_{ij} = \mu_i + \zeta_j \tag{4.3}$$

$$y_{ij} = \mu_i + \zeta_j + \boldsymbol{\lambda}_i' \boldsymbol{f}_j + e_{ij} \quad (4.4)$$

If in addition to the previous assumptions of the common factor model it is assumed that the term ζ_j is standardized and that it is uncorrelated with the error terms and with the common factors, the covariance structure implied by the RIIFA model can be written as:

$$\sum_y = \mathbf{1}\omega\mathbf{1}' + \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \quad (4.5)$$

where ω is the variance of ζ_j across all respondents.

In the RIIFA model, the parameter to be estimated is ω and not the random intercept for each examinee. To do so, it is only necessary to define an additional wording method factor in the common factor model in which all the unstandardized factor loadings are fixed to 1 (if items are not reverse coded) and ω is left free to be estimated.

Prior research has shown that wording effects can be successfully modeled through the RIIFA model, as evidenced by the improvements in model fit in comparison to models that only contain substantive factors (e.g., Abad et al., 2018; Billiet & McClendon, 2000; Kam et al., 2012; Maydeu-Olivares & Coffman, 2006; Yang et al., 2018). Besides, the RIIFA model has been shown to enhance the discriminant validity of scales (Kam et al., 2012). Despite its apparent advantages in practice, however, there is a lack of systematic studies examining the performance of the RIIFA model to estimate item and person parameters. In this regard, Savalei and Falk (2014) have evaluated its behavior to estimate item parameters when respondents make an idiosyncratic use of response scale with unidimensional structures. They found that the RIIFA model was superior to competing approaches (including the "do nothing" approach) and robust to the violation of its assumption of equal wording factor loadings across items.

4.5 Purpose of the Current Study

A principal concern regarding the use of self-report measures is the potential influence of wording effects on examinees responses. Therefore, the main motivating goal of this study was to examine the impact of different types of wording effects –carelessness, item verification

difficulty, and acquiescence– on person score estimates. In addition, we also evaluated the impact of the different types of wording effects on other parameters of interest such as model fit, factor loadings, and structural validity, for models composed of one substantive factor. Model estimates resulting from the traditional one-factor model (the "do nothing" approach) were compared to those obtained from the RIIFA model.

This study has three main unique features. First, the comprehensive evaluation of the recovery of person scores and its relationship with other parameter estimates in the presence of different wording effects. Second, the inclusion for the first time of the item verification difficulty wording effect to be examined via Monte Carlo methods. Third, the systematic evaluation of the RIIFA model to estimate person scores (and other parameters not previously studied with this model) in the presence of different wording effects.

4.6 Study 1: Impact of Carelessness on Parameter Estimation

In this study, Monte Carlo methods were employed to systematically assess the impact of the wording effect of carelessness in the performance of the 1F and RIIFA models.

4.7 Method

4.7.1 Study Design

Three independent variables were systematically manipulated: the amount of wording effect, the sample size, and the test length. These variables have been shown to affect the performance of factor analysis methods with categorical variables (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009; Garrido, Abad, & Ponsoda, 2011, 2013; Woods, 2006).

1. *Amount of wording effect* (PERC.WE): this variable indicates the percentage of items (out of the total number of items in the test) that each inconsistent examinee misresponded to. Five levels were manipulated: 10%, 20%, 30%, 40%, and 50%. The condition of absence of wording effect (PERC.WE = 0%) was included as a baseline with which to compare.

2. *Sample size* (N). Three levels were included –200, 500, and 1,000– to represent a small, a medium, and a large number of cases, respectively, for the factor analysis of categorical variables (Forero et al., 2009; Muthén & Kaplan, 1985; Savalei & Rhemtulla, 2013).

3. *Test length* (T.LENG). Three levels were included with 12, 24, and 60 items to measure the substantive construct, which may represent a short fixed-length test, a long fixed-length test, and a large item pool, respectively.

In total, the 6 × 3 × 3 (PERC.WE × N × T.LENG) factorial design produced 54 factor combinations, for each of which 100 sample replicates were generated.

4.7.2 Data Generation and Models Evaluated

Figure 4.1 presents a flow chart illustrating the main steps of the simulation study. The simulation study involved three steps: (1) generation of the uncontaminated sample data matrices, (2) generation of the sample data matrices with wording effects, and (3) estimation of the fitted models.

Step 1: Generation of uncontaminated sample data matrices. For each of the nine simulated conditions without wording effects, 100 uncontaminated (i.e., without WE) sample data matrices of symmetrically distributed categorical variables with four response options were generated. Data matrices were generated according to the bidimensional model showed in Step 1 of Figure 4.1, which contained one substantive factor representing the responses to the construct of interest and another factor representing the responses to a criterion variable (i.e., criterion variable factor, henceforth). Regarding the substantive factor, half of the items were conceptualized as PW and the other half as NW items (i.e., balanced scales). The example illustrates a 10-item (5 PW, 5 NW) test. In all the conditions, the mean substantive factor loading was fixed to .70 and loadings were randomly drawn from a uniform distribution with values ranging from .60 to .80 to generate items with variable factor loadings. Then, half of the simulated factor loadings were randomly assigned a negative sign to simulate the factor

loadings of the NW items. Additionally, the criterion variable factor was simulated by generating the responses to an item with a standardized loading of 1.0 on such factor. The substantive factor and the criterion variable correlated strongly (r = .50) according to Cohen (1988).

Sample data matrices were simulated according to the common factor model procedure described next. First, the reproduced population correlation matrix (with communalities in the diagonal) was computed:

$$\mathbf{R_R} = \mathbf{\Lambda \Phi \Lambda'} \tag{4.6}$$

where $\mathbf{R_R}$ is the reproduced population correlation matrix, *lambda* ($\mathbf{\Lambda}$) is the measurement model (i.e., a $k \times 2$ factor loading matrix for k variables and 2 factors – the substantive factor and the criterion variable factor–) and *phi* ($\mathbf{\Phi}$) is the structure matrix of the latent variables (i.e., a 2×2 matrix of correlations among the substantive factor and the criterion variable factor).

The population correlation matrix $\mathbf{R_P}$ was then obtained by inserting unities in the diagonal of $\mathbf{R_R}$, thereby raising the matrix to full rank. The next step was performing a Cholesky decomposition of $\mathbf{R_P}$, such that:

$$\mathbf{R_P} = \mathbf{U'U} \tag{4.7}$$

Subsequently, the sample data matrix of continuous variables was computed as:

$$\mathbf{X} = \mathbf{ZU} \tag{4.8}$$

where $\mathbf{Z}$ is a matrix of random standard normal deviates with rows equal to the sample size and columns equal to the number of variables.

The resulting continuous variables were categorized (except the criterion variable, which was not included in the following steps) by applying the following threshold values so that they had symmetrical distributions: -1.5, 0, and 1.5 (Garrido et al., 2011, 2013).

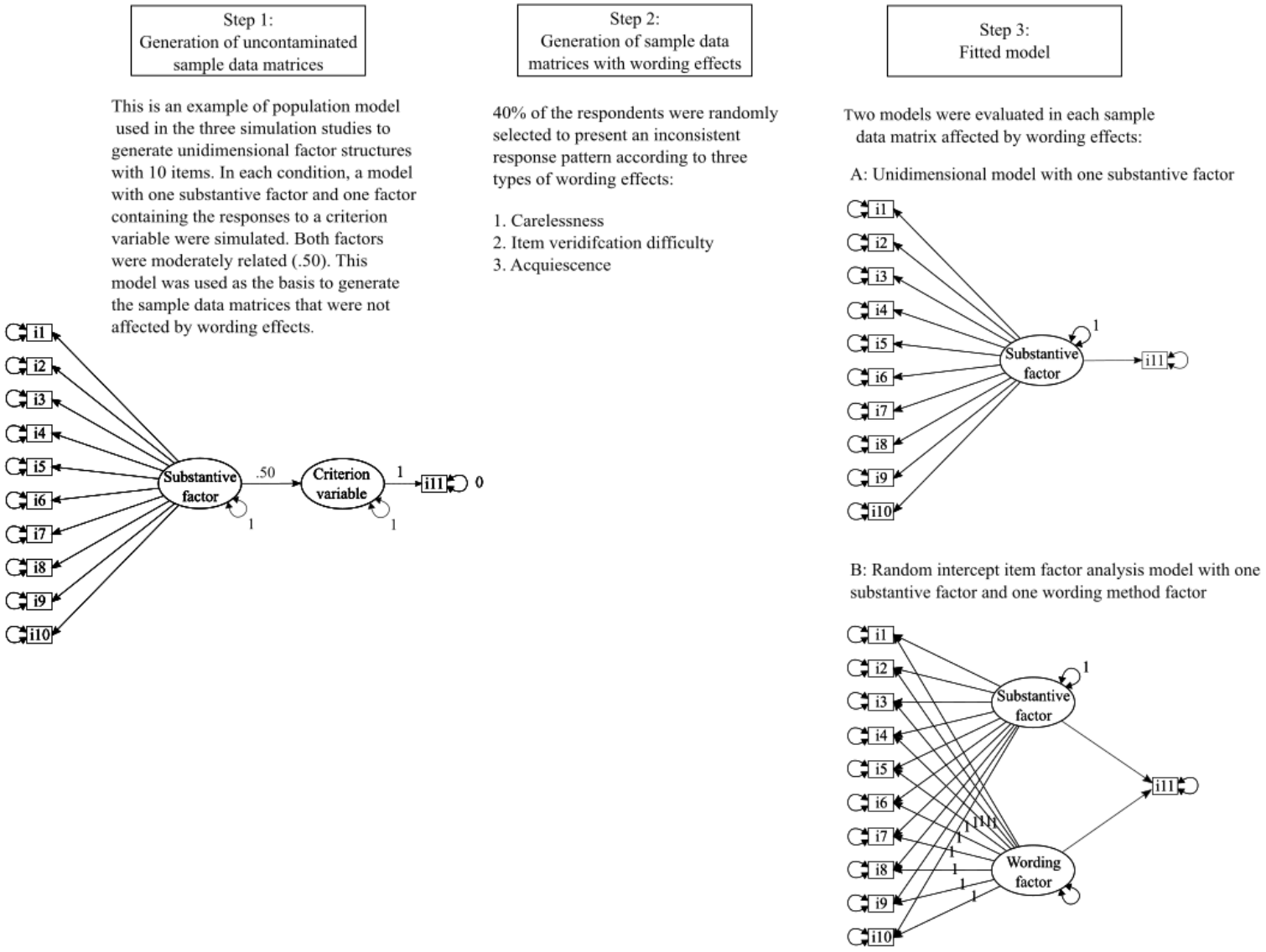

Figure 4.1. Flow chart diagram describing the main steps of the three simulation studies.

Step 2: Generation of sample data matrices with wording effects. Wording effects were generated by introducing inconsistent responses in the uncontaminated sample data matrices previously simulated (Step 2 in Figure 4.1). To do so, 40% of the simulees were randomly selected to present inconsistent responses. Then, following previous research (Schmitt & Stults; Woods, 2006), carelessness response patterns were simulated by reversing the answers to NW items (1 = 4, 2 = 3, 3 = 2, and 4 = 1) for each inconsistent respondent according to the desired amount of wording effects in each case. For each uncontaminated sample data matrix, five sample data matrices with wording effects were generated according to the five levels of PERC.WE established. Table 4.1 includes some examples of response patterns for hypothetical examinees responding carelessly to NW items. These "respondents" were postulated to respond inconsistently to 50% of the items of a 10-item test. We decided to simulate carelessness to NW items arbitrarily, based on previous studies. This would follow a real-life scenario where the first items of a test were PW and/or the examinee had an idea that the trait being measured had a positive valence. Nevertheless, simulating carelessness to the PW items would have yielded the same general conclusions.

Step 3: Estimation of the fitted models. The two structural equation models represented in Step 3 of Figure 4.1 were estimated for each of the simulated sample data matrices. The first model (Figure 4.1, Step 3, A) had one substantive factor measured by the simulated target (categorical) items, and the (continuous) item representing the observed scores for the simulees on a criterion variable that was regressed on the substantive factor. As the main core of this model is the traditional one-factor model with a substantive factor, we will refer to this model as 1F. The second model (Figure 4.1, Step 3, B) included the RIIFA approach to model one substantive factor and one method factor to control for wording effects (Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006), and the observed criterion variable that was regressed on the substantive factor and the wording factor. In the RIIFA

model, the loadings in the wording factor were fixed to 1 because sample data matrices contained unrecoded item scores, and the variance of the wording method factor was estimated. As the main core of this second model is the estimation of the RIIFA, we will refer to it as RIIFA. Both models were estimated using robust weighted least squares estimation based on a matrix of polychoric correlations (WLSVM, see Muthén & Muthén, 1998–2012).

4.7.3 Assessment Criteria

Although the primary estimates of interest were the substantive factor scores, the performance of each model was also evaluated according to three other fundamental aspects in model validation: model fit, recovery of substantive factor loadings, and structural validity. For each model, the following assessment criteria were obtained:

Model fit: It was evaluated according to the root mean square error of approximation (RMSEA) and the comparative fit index (CFI). For the CFI, values of .90 or greater indicate acceptable fit and values of .95 or greater represent good fit, whereas RMSEA values between .05 and .08 are indicative of acceptable model fit and values below .05 represent good fit (Hu & Bentler, 1999; McDonald & Ho, 2002).

Recovery of the substantive factor loadings: It was evaluated for PW and NW items separately by computing the mean bias error (MBE) and the mean absolute error (MAE) in each case:

$$MBE = \frac{\sum(\hat{\lambda} - \lambda)}{k}, \tag{4.9}$$

$$MAE = \frac{\sum|(\hat{\lambda} - \lambda)|}{k} \tag{4.10}$$

where k is the number of PW or NW items, $\hat{\lambda}$ is the estimated loading on the substantive factor, and λ is the true loading on the substantive factor.

An MBE of 0 reflects a total lack of bias, whereas negative and positive MBE values indicate that loadings were underestimated and overestimated in absolute value, respectively, for PW items, and the opposite for NW items. For the MAE, higher values signal larger biases in estimating the true factor loadings, while a value of 0 indicates that the factor loadings are estimated with perfect accuracy.

Recovery of the substantive scores: It was evaluated with the correlation between the uncontaminated substantive factor scores (estimated by applying the 1F model to the sample data matrix without wording effects) and the contaminated factor scores that were estimated by applying each model to the data matrix with wording effects. A mixed analysis of variance (ANOVA) was performed in order to evaluate the differences between the 1F and the RIIFA models in the recovery of the uncontaminated substantive scores across the manipulated conditions. The dependent variable was the total recovery of the substantive factor scores, the repeated measures within-subjects independent variable was the model (1F, RIIFA), and the between-subjects independent variables were the amount of wording effects, the sample size, and the test-length. Due to the low convergence rate of the RIIFA model with 10% of wording effects across the three studies, the cases for that condition were not included in these analyses. Only those higher order interactions with large or near-large effects sizes were interpreted. According to Cohen (1988), partial eta squared ($\eta_p{}^2$) values of .01, .06, and .14 or greater represent small, medium, and large effects, respectively.

Structural validity: It was evaluated through the magnitude of the regression coefficients associated to the substantive factor and the wording factor (only for the RIIFA model), as well as the proportion of variance explained by the model (R^2).

The programs used to run the simulation were developed with M*plus* 7 (Muthén & Muthén, 1998-2012) and the R package *MplusAutomation* (Hallquist & Wiley, 2018). The

statistical analyses and the simulation were performed using SPSS (v.23) and R (R Core Team, 2018), respectively.

4.8 Results

4.8.1 Convergence Rates

The convergence rates reported in this section indicate for each model tested (1F, RIIFA) the proportion of estimated solutions that produced simultaneously the fit statistics, the matrix of factor loadings, the factor scores, the regression coefficients, and the R^2. The convergence rate of the 1F model was always 100%. The overall convergence rate for the RIIFA model was 92.71%. Nonconvergence occurred with low amounts of wording effects (10% or 20%) and the tests with 10 or 20 items. In those conditions, convergence rates improved with larger tests and higher amounts of wording effects: 26.67% (PERC.WE = 10%, 10 items), 71.00% (PERC.WE = 10%, 20 items), and 93.00% (PERC.WE = 20%, 10 items).

4.8.2 Model Fit

Panel A of Figure 4.2 shows the CFI and RMSEA values obtained with both models through all the simulated sample data matrices across different amounts of wording effects. With lower amounts of wording effects (particularly 10%), both models showed excellent model fit presenting always the CFI and RMSEA mean values ($\overline{CFI}$, $\overline{RMSEA}$) very close to 1 and 0, respectively. However, as the amount of wording effect increased, the differences between models were more notable: the 1F model gradually presented poorer fit and the values of the fit indices progressively departed from acceptable fit, reaching the worst values with PERC.WE = 50% ($\overline{CFI}$ = .51, $\overline{RMSEA}$ = .14). In contrast, the RIIFA model showed almost perfect model fit with any amount of wording effects.

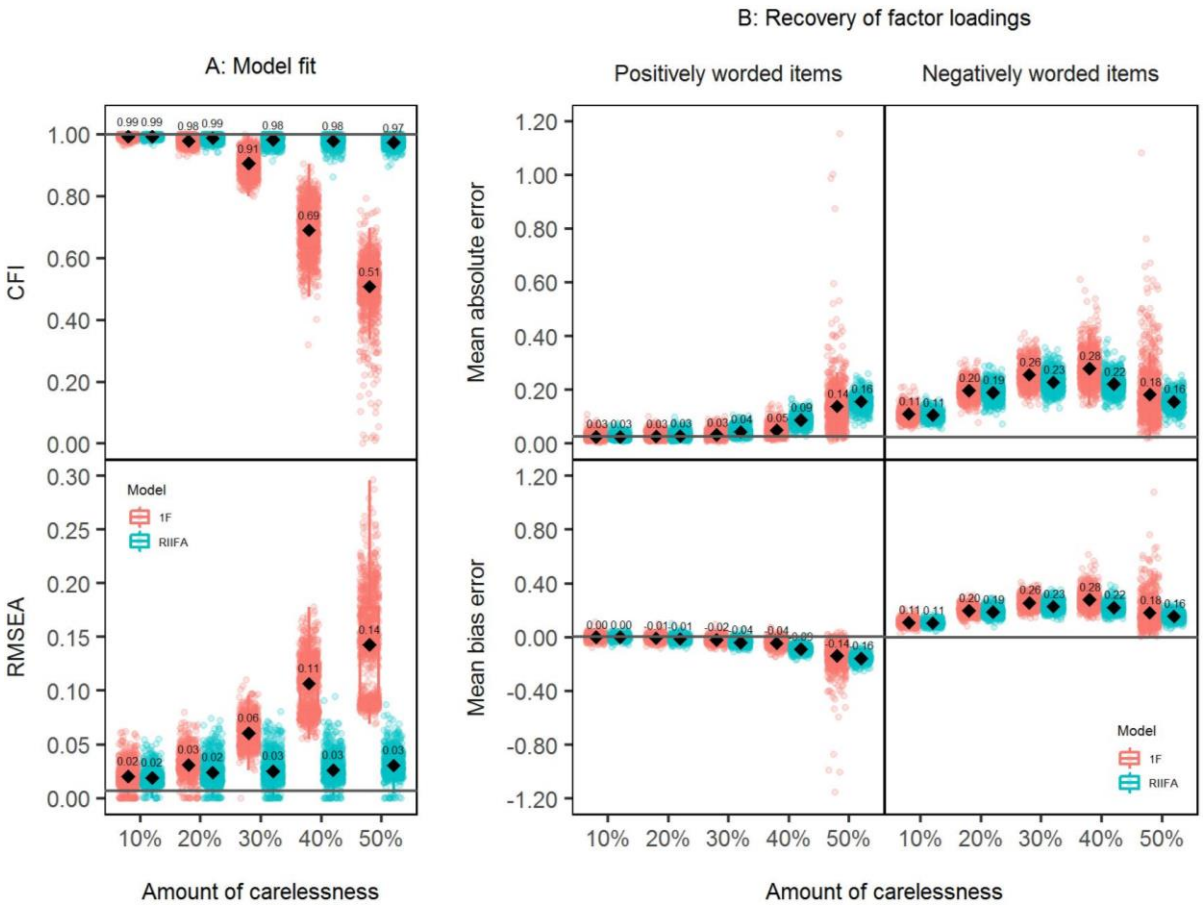

Figure 4.2. Model fit and recovery of substantive factor loadings for positively worded and negatively worded items with the 1F model and the RIIFA model in the presence of carelessness. 1F = unidimensional model with one substantive factor; RIIFA = random intercept item factor analysis model with one substantive factor and one wording method factor. In panel A, the horizontal grey lines represent the mean CFI and RMSEA values for the condition with 0% of wording effect. In panel B, the horizontal grey lines represent the mean absolute error and the mean bias error for the condition with 0% of wording effect.

4.8.3 Recovery of the Substantive Factor Loadings

Panel B of Figure 4.2 shows the individual MBEs and MAEs for the simulated sample data matrices obtained with both models (1F, RIIFA) for each type of item (PW, NW) across the different amounts of wording effects. Looking at the average MAE ($\overline{\text{MAE}}$) values, in general both models produced less accurate estimations for NW than for PW items across conditions, except with 50% of misresponded items where both models performed similarly and also produced similar estimates between them (e.g., for NW items, $\overline{\text{MAE}}$[1F] = .18 and $\overline{\text{MAE}}$[RIIFA] = .16) and for both types of items (e.g., for the RIIFA model, $\overline{\text{MAE}}$[PW] ≈

$\overline{\text{MAE}}$[NW] = .16). Moreover, estimates with both models were gradually less precise as the percentage of misresponded items increased. The only exception was found for NW items when the amount of wording effect grew from 40% to 50%: in this condition, the $\overline{\text{MAE}}$ decreased markedly from .28 to .18 for the 1F model, and from .22 to .16 for the RIIFA model. Besides, a look at the MBE values revealed that both models tended to underestimate the factor loadings of any type of item, and that this tendency increased with higher amounts of wording effect.

4.8.4 Recovery of the Substantive Factor Scores

In order to better understand the performance of both models, the recovery was evaluated by computing the correlation between the uncontaminated and contaminated factor scores in three ways for each simulated sample data matrix: (a) considering the scores for all the respondents (consistent and inconsistent; henceforth 'Total recovery'), (b) considering consistent respondents scores exclusively (henceforth 'Recovery for consistent respondents'), and (c) considering inconsistent respondents scores exclusively (henceforth 'Recovery for inconsistent respondents'). Results from the mixed ANOVA (Table 4.2) revealed that a large effect size ($\eta_p{}^2$ = .17, p < .001) was associated to the differences in performance between models in favor of the 1F model which showed to be slightly superior to the RIIFA across conditions (overall, the mean correlation, $\bar{r}$, was .93 and .90, respectively, for the 1F and RIIFA model, respectively). Almost all interactions displayed $\eta_p{}^2$ values lower or equal to .02. Only the two-way interaction Model × Amount of wording effects showed a near-large effect size ($\eta_p{}^2$ = .11, p < .001). This interaction is depicted in the upper section of panel A in Figure 4.3, and shows that both models performed almost similarly with 20% of carelessness, whereas with 50% of carelessness the 1F model ($\bar{r}$ = .78) proved to be slightly superior to the RIIFA ($\bar{r}$ = .76) in recovering the uncontaminated person scores. The performance of both models gradually deteriorated as the amount of wording effect increased.

Table 4.2

Mixed Analysis of Variance Effect Sizes for the Wording Effects of Carelessness, Item Verification Difficulty, and Acquiescence

Effect type/Variables	Carelessness	Item verification difficulty	Acquiescence
Main effect			
Model	**.17*** **	**.14*** **	**.20*** **
Two-way interactions			
Model × Amount of wording effects	.11***	.05***	**.24*** **
Model × Sample size	.00**	.00*	.02***
Model × Test length	.00	.02***	.00
Three-way interactions			
Model × Amount of wording effects × Sample size	.01***	.01***	.04***
Model × Amount of wording effects × Test length	.02***	.02***	.01***
Model × Sample size × Test length	.00	.00*	.00
Four-way interaction			
Model × Amount of wording effects × Sample size × Test length	.00	.00	.00

Note. Tabled values are partial eta squared (η_p^2) estimates of variance explained by each of the effects shown. The dependent variable was the correlation between the uncontaminated substantive scores and the contaminated substantive scores. Large effect sizes (η_p^2 = .14) are bolded and underlined. *: p<.05, **: p<.01, ***p<.001

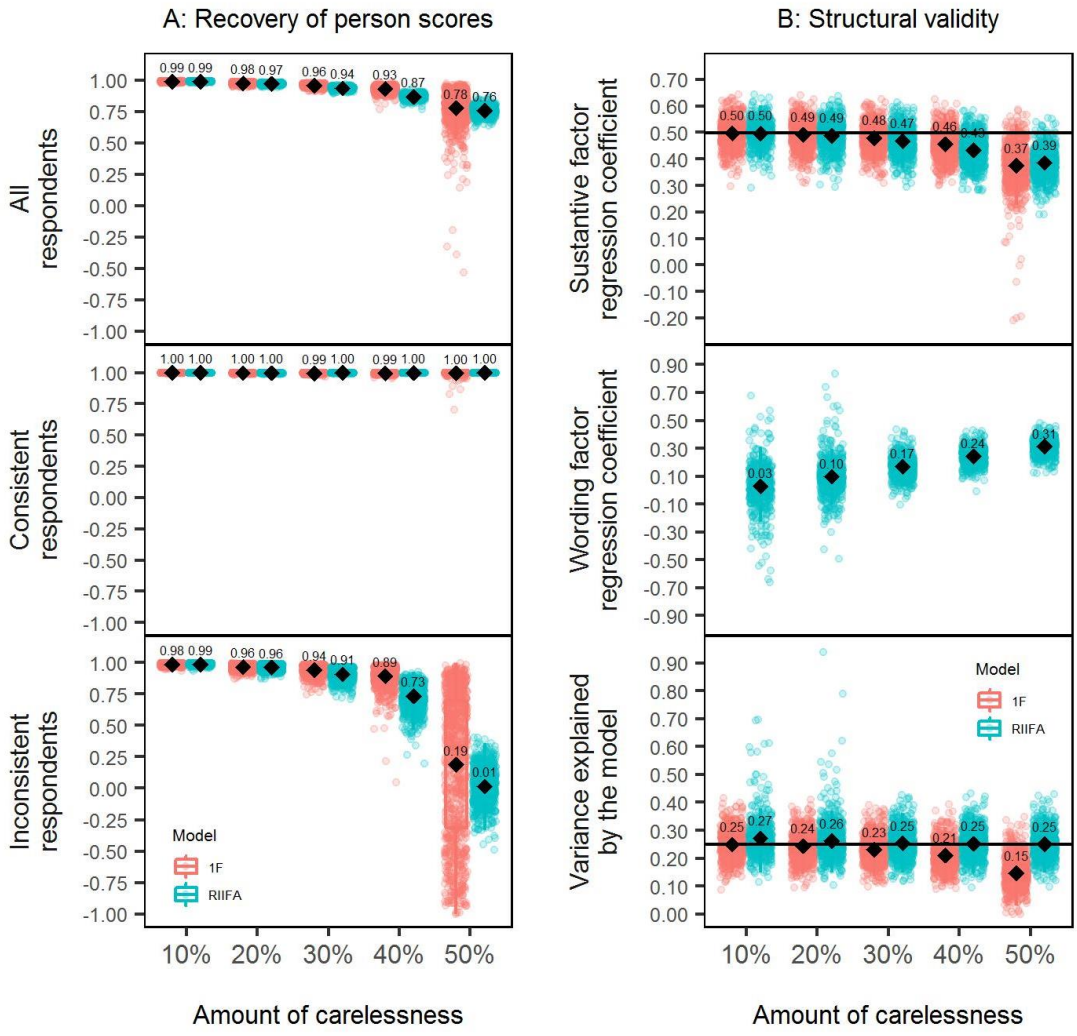

Figure 4.3. Recovery of substantive factor scores and estimation of structural validity with the 1F model and the RIIFA model in the presence of carelessness. In panel A, represented values are Pearson correlations, and results of the recovery of uncontaminated scores are presented for all the respondents, separately for consistent respondents, and separately for inconsistent respondents. 1F = unidimensional model with one substantive factor; RIIFA = random intercept item factor analysis model with one substantive factor and one wording method factor. In panel B, the horizontal grey lines represent the mean substantive factor regression coefficient and the mean variance explained by the model with 0% of wording effect.

The middle and lower sections of panel A in Figure 4.3 show the recovery of substantive factor scores for the consistent and inconsistent respondents, respectively, across the two-way interaction Model × Amount of wording effect. Regarding the consistent respondents, both models always estimated with perfect accuracy the substantive scores of the consistent

respondents (the mean correlation between the uncontaminated and contaminated scores was always 1.00). Looking at the results for the inconsistent respondents, the patterns showed by both models mirrored the results previously described for the total recovery but the $\bar{r}$ values were systematically lower. It should be noted that the recovery for these respondents with 50% of wording effects was especially poor if looking at the average ($\bar{r}$[1F] = .19, $\bar{r}$[RIIFA] = .01). This might be explained because if a person misresponds to all the items in one direction (all PW or all NW) it is impossible to recover his/her uncontaminated score because there is no way to know if the correct score is what he/she responded to PW items or what he/she answered to the NW items. In other words, the answers of this person to both types of items are equally consistent.

In order to better understand the previous results, Figure 4.4 shows for each model a series of scatter-plots to illustrate the relationship between the uncontaminated and contaminated substantive scores as the amount of wording effects increased. To do so, we simulated a sample data matrix with 1000 respondents and 20 items, which was later modified according to the levels of carelessness established. As shown before, for consistent respondents (colored in black) the substantive scores are always estimated with total precision with both models because they delineate a perfect diagonal straight line. In the case of inconsistent respondents (colored in red), with both models the contaminated scores for respondents that had low uncontaminated scores tend to be increasingly biased upward, whereas the contaminated scores for respondents with high uncontaminated scores are progressively biased downward. This displacement is progressively more noticeable as the percentage of misresponded items is higher. Figure 4.4 also presents the correlation between the uncontaminated substantive scores and the estimated wording factor scores for the RIIFA model. Overall, consistent respondents had wording scores of medium magnitude (i.e., around 0) independently of the value of their uncontaminated substantive score. Regarding

inconsistent respondents, the estimated wording scores were increasingly correlated in a positive way with the substantive scores as the amount of wording effect was greater. This means that wording scores increasingly reflect the uncontaminated trait level of these examinees as more items are answered inconsistently.

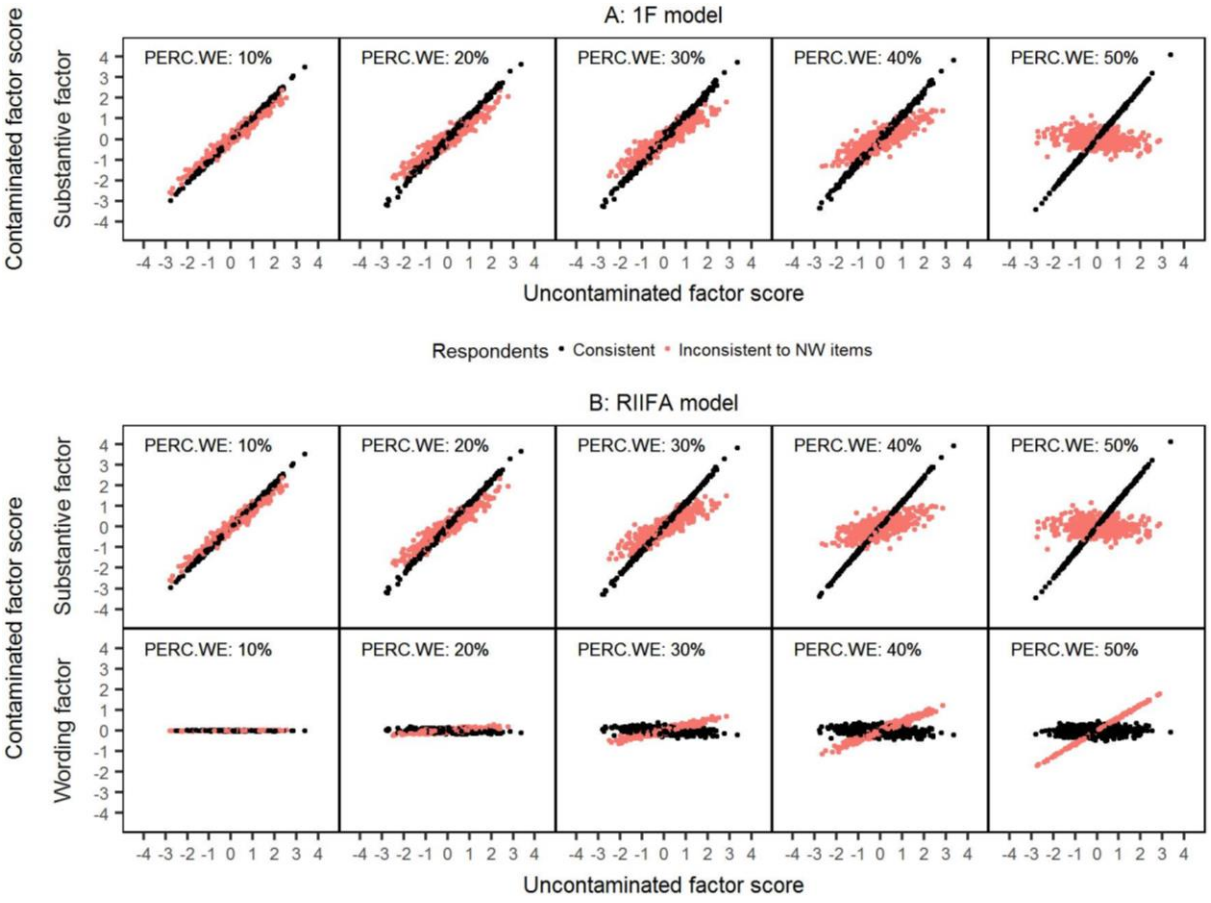

Figure 4.4. Example of recovery of the substantive factor scores with the 1F and RIIFA models in the presence of carelessness. The data represented corresponds to simulated unidimensional data sets with 1000 respondents and 20 variables. 1F = unidimensional model with one substantive factor; RIIFA = random intercept item factor analysis model with one substantive factor and one wording method factor; PERC.WE = amount of wording effect; NW = negatively worded.

4.8.5 Structural Validity

The panel B of Figure 4.3 shows the regression coefficients associated to the substantive factor for both models and the wording factor of the RIIFA, as well as the proportion of explained variance by each model (R^2). In terms of the mean regression coefficient for the substantive factor, both models showed a tendency to produce downwardly biased estimates, on average, with higher amounts of wording effect. In addition, the regression coefficient associated to the wording factor showed a tendency to increase gradually as the amount of carelessness grew, reaching non-negligible values in the conditions of greater carelessness. This might explain why the mean proportion of variance explained by the model was moderately greater for the RIIFA model in comparison to the 1F model. Indeed, it should be noted that the RIIFA model always reproduced the same amount of variance (on average) as the model fitted in the dataset without wording effects.

4.9 Study 2: Impact of Item Verification Difficulty on Parameter Estimation

In this study, Monte Carlo methods were employed to systematically assess the impact of the wording effect of item verification difficulty in the performance of the 1F and RIIFA models.

4.10 Method

The study design and the procedure followed to generate the uncontaminated sample data matrices (see Figure 4.1) was the same as the one described in Study 1. However, in this case inconsistent respondents were simulated by reversing the answers (1 = 4, 2 = 3, 3 = 2, and 4 = 1) to PW items if the uncontaminated substantive score of a respondent was below 0, or to NW items if the uncontaminated substantive score of a respondent was above 0. That is, it was assumed that a person responded correctly to true affirmations and responded incorrectly to false affirmations. When the PERC.WE was below 50%, item responses were randomly selected and reversed until the desired amount of wording effects (% of items answered

inconsistently) for each respondent was reached. The proportion of respondents in each database that answered inconsistently was again fixed at 40%. Finally, the same assessment criteria of Study 1 were obtained to evaluate the performance of both models.

4.11 Results

4.11.1 Convergence Rates

As in Study 1, only the RIIFA model showed nonconvergence solutions, with an overall convergence rate of 93.38%. The pattern of convergence rates was similar to that found in Study 1: nonconvergence occurred with low amounts of wording effects (10% or 20%) and the tests with 10 or 20 items. Convergence rates improved with larger tests and higher amounts of wording effects: 30.33% (PERC.WE = 10%, 10 items), 93.67% (PERC.WE = 10%, 20 items), and 77.00% (PERC.WE = 20%, 10 items).

4.11.2 Model Fit

Panel A of Figure 4.5 shows the CFI and RMSEA values obtained with both models through all the simulated sample data matrices across different amounts of wording effect. Results revealed a similar trend to that found in Study 1: both models performed similarly, showing perfect fit, with lower amounts of item verification difficulty. However, as the amount of wording effect increased, the differences were more notable in favor of the RIIFA model, which consistently showed almost perfect model fit, whereas the 1F model gradually presented poorer fit.

4.11.3 Recovery of the Substantive Factor Loadings

Panel B of Figure 4.5 shows the individual MBEs and MAEs for the simulated sample data matrices obtained with both models (1F, RIIFA) for each type of item (PW, NW) across the different amounts of wording effects. Looking at the $\overline{\text{MAE}}$, in general both models produced similar estimates between them and for both types of items for any amount of wording effects. As in Study 1, in general the estimates with both models were gradually less precise as the

percentage of misresponded items increased, except for the 50% PERC.WE condition which was similar to the 40% condition. As in Study 1, the MBE values showed that both models tended to underestimate the factor loadings of any type of item, and that this tendency increased with higher amounts of wording effects.

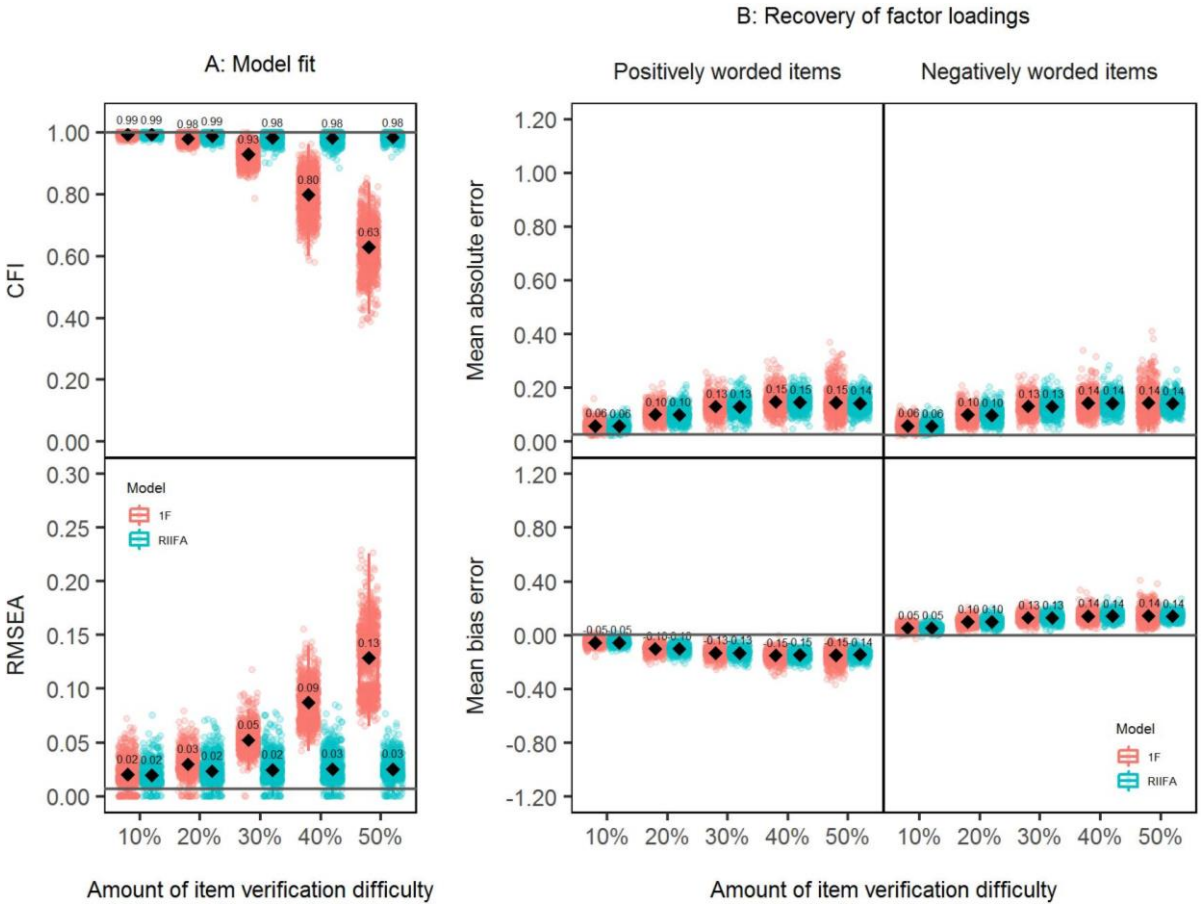

Figure 4.5. Model fit and recovery of substantive factor loadings for positively worded and negatively worded items with the 1F model and the RIIFA model in the presence of item verification difficulty. 1F = unidimensional model with one substantive factor; RIIFA = random intercept item factor analysis model with one substantive factor and one wording method factor. In panel A, the horizontal grey lines represent the mean CFI and RMSEA values for the condition with 0% of wording effect. In panel B, the horizontal grey lines represent the mean absolute error and the mean bias error for the condition with 0% of wording effect.

4.11.4 Recovery of the Substantive Factor Scores

The results of the mixed ANOVAs comparing the precision of the factor score estimates across the manipulated conditions for the 1F and RIIFA models showed that although a large effects size (η_p^2 = .14, p < .001) was associated to the differences in performance between models, it had no practical relevance because, on average, the overall recovery was similar for both models (.90 for the 1F and .89 for the RIIFA). This large effect size emerged because of the low variability of the individual results for the replications in the simulation (Pek & Flora, 2018). Almost all interactions displayed η_p^2 values lower or equal to .02, except the two-way interaction Model × Amount of wording effect, which had a larger but still small effect size (η_p^2 = .05, p < .001). Panel A in Figure 4.6 displays this interaction which has a similar trend to the one describe in Study 1: both models performed similarly with 20% of item verification difficulty, but they tended to display some negligible differences in favor of the 1F model: the maximum difference that both models showed regarding the mean total recovery was of .01, which is negligible, with the greatest amount of item verification difficulty). The performance of both models gradually deteriorated as the amount of wording effect increased.

The middle and lower sections of panel A in Figure 4.6 show the recovery of the substantive factor scores for consistent and inconsistent respondents, respectively, across the two-way interaction of Model × Amount of wording effects. These results mirrored those obtained in Study 1 for the wording effect of carelessness, with the recovering being approximately perfect for the consistent respondents and increasingly poorer with greater wording effects for the inconsistent respondents.

As in Study 1, the results from a simulated sample data matrix with 1000 respondents and 20 items was used to obtain a series of scatter-plots to illustrate the relationship between the uncontaminated and contaminated substantive scores across different amounts of wording effects (see Figure 4.7). The trends observed for consistent respondents (colored in black) and

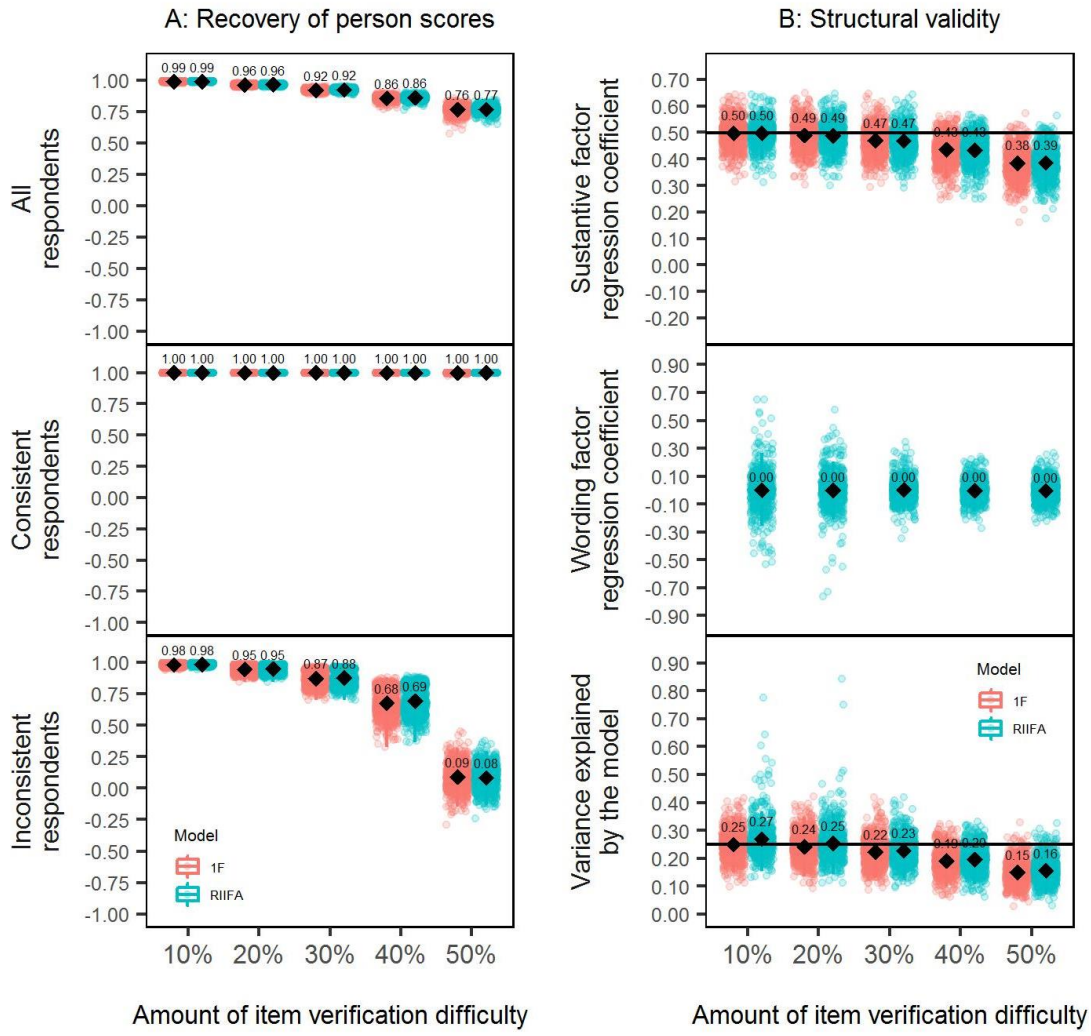

Figure 4.6. Recovery of substantive factor scores and estimation of structural validity with the 1F model and the RIIFA model in the presence of item verification difficulty. In panel A, represented values are Pearson correlations, and results of the recovery of uncontaminated scores are presented for all the respondents, separately for consistent respondents, and separately for inconsistent respondents. 1F = unidimensional model with one substantive factor; RIIFA = random intercept item factor analysis model with one substantive factor and one wording method factor. In panel B, the horizontal grey lines represent the mean substantive factor regression coefficient and the mean variance explained by the model with 0% of wording effect.

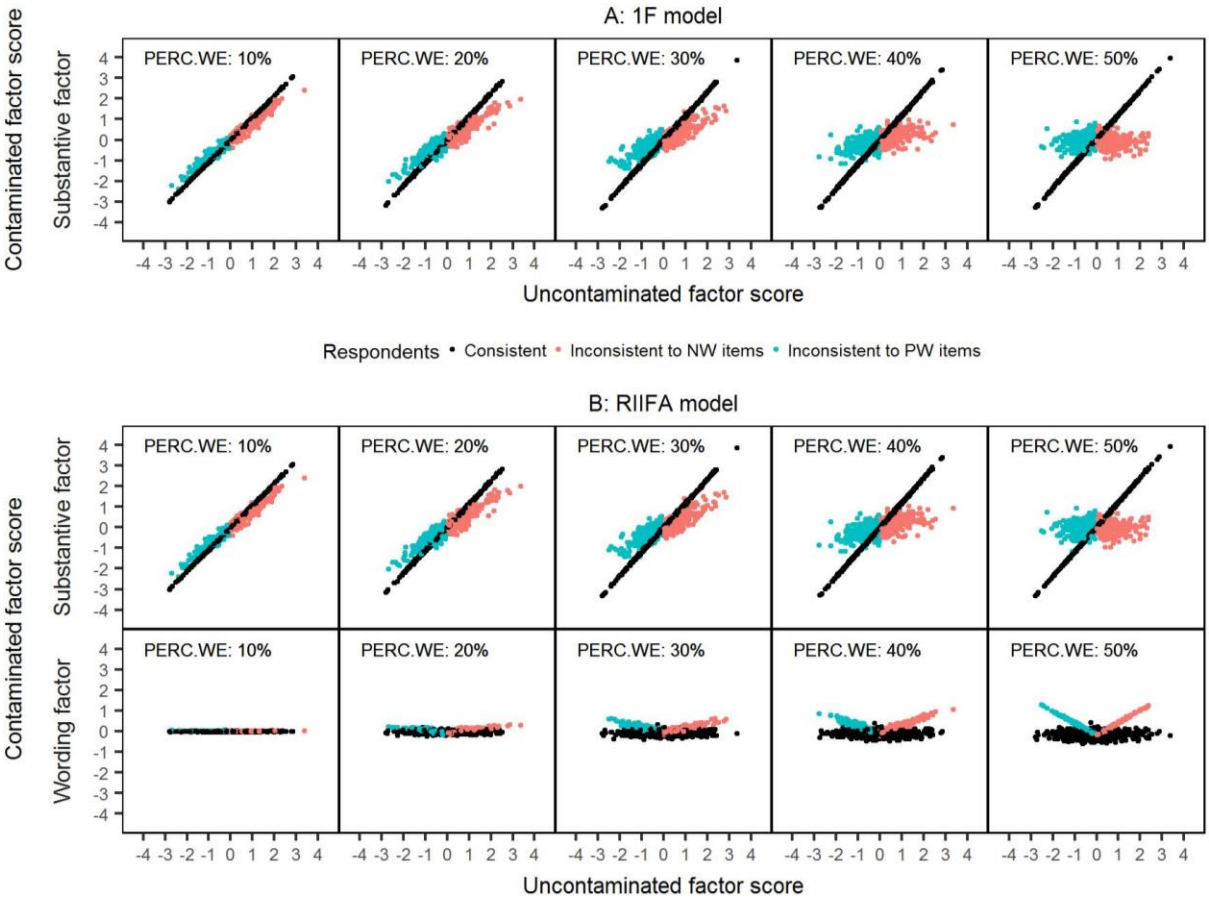

Figure 4.7. Example of recovery of the substantive factor scores with the 1F and RIIFA models in the presence of item verification difficulty. The data represented corresponds to simulated unidimensional data sets with 1000 respondents and 20 variables. 1F = unidimensional model with one substantive factor; RIIFA = random intercept item factor analysis model with one substantive factor and one wording method factor; PERC.WE = amount of wording effect; NW = negatively worded; PW = positively worded.

inconsistent respondents (colored in blue or red depending on whether they misresponded to PW or NW items, respectively) with both models mirrored the ones found in the case of carelessness (see Study 1). The correlation between the uncontaminated substantive scores and the estimated wording factor scores was also represented for the RIIFA model. For examinees who misresponds to NW items, wording scores relates positively with the uncontaminated substantive score, and this relation is stronger as the amount of item verification difficulty is greater. Contrarily, for examinees who misresponds to PW items, wording scores relate

inversely to the uncontaminated substantive score, and the magnitude of such correlation will be higher with higher amounts of wording effect.

4.11.5 Structural Validity

Panel B of Figure 4.6 shows the regression coefficients associated to the substantive factor for both models and the wording factor of the RIIFA, as well as the proportion of explained variance by each model (R^2). In terms of the mean regression coefficient for the substantive factor, both models showed a tendency to produce downwardly biased estimates, on average, with higher amounts of wording effects. Additionally, the regression coefficient associated to the wording factor had a mean of zero across conditions. Both models tended to underestimate the proportion of variance as the amount of wording effects increased, and although the RIIFA was slightly superior to the 1F model, the gains in variance explained were minimal.

4.12 Study 3: Impact of Acquiescence on Parameter Estimation

In this study Monte Carlo methods were employed to systematically assess the impact of the wording effect of acquiescence in the performance of the 1F and RIIFA models.

4.13 Method

The study design and the procedure to generate the uncontaminated sample data matrices (see Figure 4.1) were similar to those described previously in Studies 1 and 2. In this case, to simulate acquiescent respondents we assumed that these individuals would select fewer response categories implying higher levels of disagreement (1 and 2) than response options representing higher levels of agreement (3 and 4). Thus, for the inconsistent respondents we arbitrarily assigned to each response category a different probability of being changed so that inconsistent respondents were generated by switching more answers with 1 than answers with 2, and more answers with 2 than answers with 3. The answers with 4 were not modified (this response option implied the highest level of agreement). The probabilities of being selected for

change for response categories 1, 2, and 3 were .50, .33, and .17, respectively. Once a response category was selected to be changed for an inconsistent respondent, its values were modified in the following manner: 1 = 3, 2 = 3 or 4 (being the two values equally likely), and 3 = 4. Item responses were changed for each inconsistent simulee until reach the corresponding amount of wording effect. Acquiescent respondents were selected using the *sample*() function from the base R package (R Core Team, 2018). Finally, the same assessment criteria of the two prior studies were used to evaluate the performance of both models.

4.14 Results

4.14.1 Convergence Rates

As in Studies 1 and 2, only the RIIFA model produced solutions that did not converge, with an overall convergence rate of 94.82%. The pattern of convergence rates was similar to that found in Studies 1 and 2: nonconvergence occurred with low amounts of wording effects (10% or 20%) and tests with 10 or 20 items. Convergence rates improved with larger tests and higher amounts of wording effects: 31.00% (PERC.WE = 10%, 10 items), 93.67% (PERC.WE = 10%, 20 items), and 97.67% (PERC.WE = 20%, 10 items).

4.14.2 Model Fit

Panel A in Figure 4.8 shows the CFI and RMSEA values obtained with both models through all the simulated sample data matrices across different amounts of wording effects. Results were similar to those found in Studies 1 and 2: both models performed similarly, showing perfect fit, with lower amounts of acquiescence. However, as the amount of wording effect increased, the differences were more notable in favor of the RIIFA model, which consistently showed almost perfect model fit, whereas the 1F model gradually presented poorer fit.

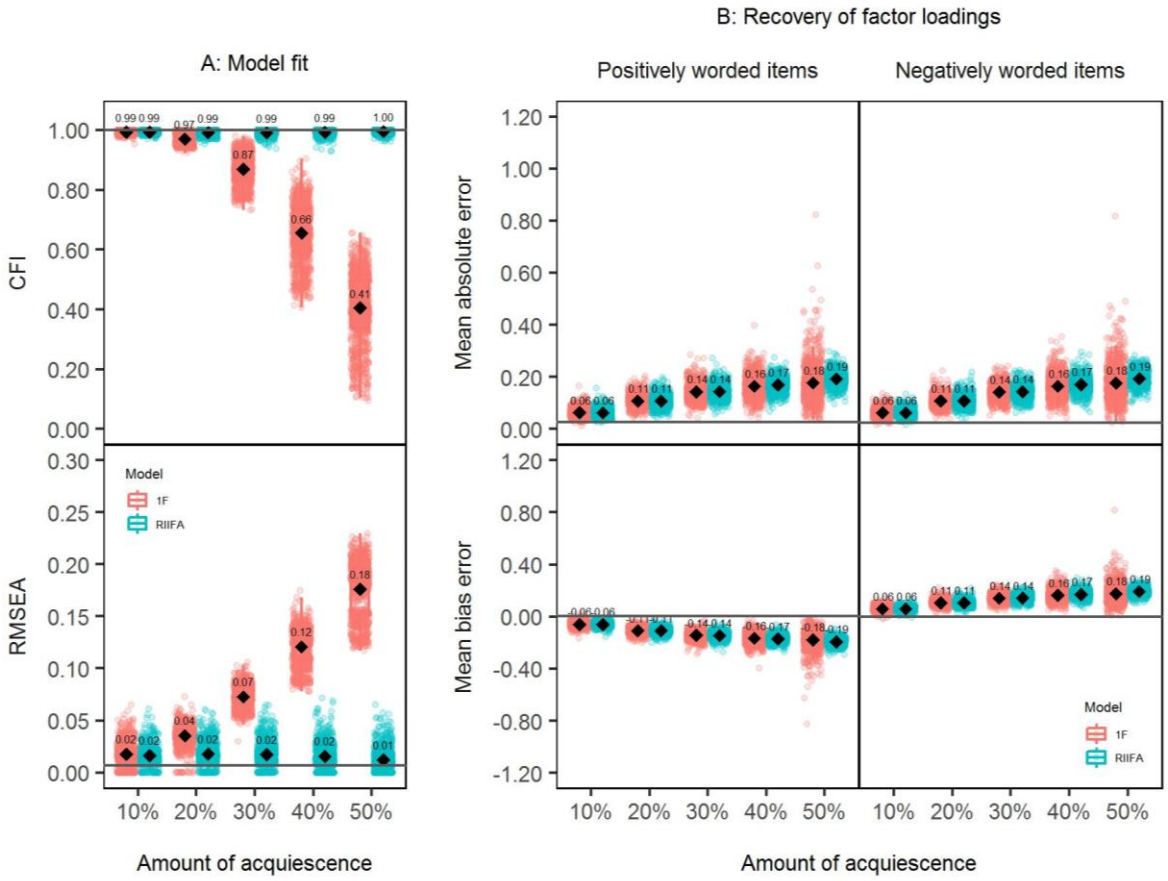

Figure 4.8. Model fit and recovery of substantive factor loadings for positively worded and negatively worded items with the 1F model and the RIIFA model in the presence of acquiescence. 1F = unidimensional model with one substantive factor; RIIFA = random intercept item factor analysis model with one substantive factor and one wording method factor. In panel A, the horizontal grey lines represent the mean CFI and RMSEA values for the condition with 0% of wording effect. In panel B, the horizontal grey lines represent the mean absolute error and the mean bias error on average for the condition with 0% of wording effect.

4.14.3 Recovery of the Substantive Factor Loadings

Panel B in Figure 4.8 shows the individual MBEs and MAEs for the simulated sample data matrices obtained with both models (1F, RIIFA) for each type of item (PW, NW) across the different amounts of wording effect. Looking at the $\overline{\text{MAE}}$, in general both models produced similar estimates between them and for both types of items with any amount of wording effect, except with 40% or more of wording effect where the RIIFA model was slightly less accurate than the 1F. As in Studies 1 and 2, in general estimates with both models were gradually less

precise with higher percentages of misresponded items. As in previous studies, the mean MBE values showed that both models tended to underestimate the factor loadings of any type of item, and that this tendency increased with higher amounts of wording effects.

4.14.4 Recovery of the Substantive Factor Scores

To evaluate the differences between models, a mixed ANOVA was performed with the same specifications as in Studies 1 and 2. Similarly to Study 2, a large effects size (η_p^2 = .20, p < .001) was associated to the differences in performance between models but it had no practical relevance because, on average, the overall recovery was similar for both models (.95 in both cases). This large effect size emerged because of the low variability of the individual results for all the replications in the simulation (Pek & Flora, 2018). Only the two-way interaction Model × Amount of wording effect reached a large effect size (η_p^2 = .24, p < .001), whereas the remaining interactions had η_p^2 values lower or equal to .04. The upper section of panel A in Figure 4.9 displays this interaction which has a similar trend to the one described in prior studies: both models performed similarly with 20% of acquiescence, and they displayed very small differences (.01) in favor of the 1F model with 50% of acquiescence. This differences had no practical relevance, and the large effect size emerged because of the low variability of the individual results for the replications in each condition represented in panel A of Figure 4.9 (Pek & Flora, 2018). The performance of both models gradually deteriorated as the amount of wording effect increased.

The middle and lower sections of panel A in Figure 4.9 show the recovery of the substantive factor scores for consistent and inconsistent respondents, respectively, across the two-way interaction Model × Amount of wording effect. Results for both consistent and inconsistent respondents are similar to those described in Studies 1 and 2 for the wording effects of Carelessness and Item verification difficulty, respectively.

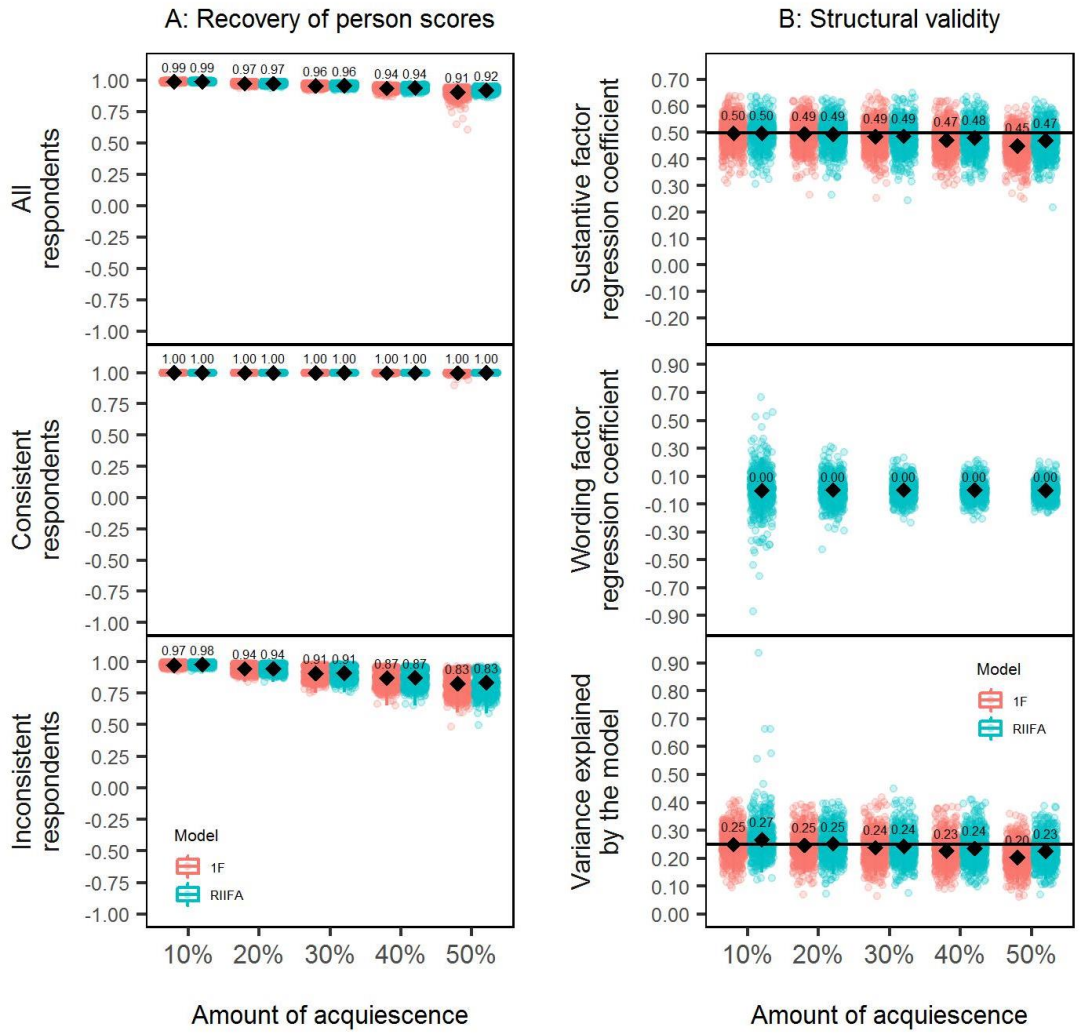

Figure 4.9. Recovery of substantive factor scores and estimation of structural validity with the 1F model and the RIIFA model in the presence of acquiescence. In panel A, represented values are Pearson correlations, and results of the recovery of uncontaminated scores are presented for all the respondents, separately for consistent respondents, and separately for inconsistent respondents. 1F = unidimensional model with one substantive factor; RIIFA = random intercept item factor analysis model with one substantive factor and one wording method factor. In panel B, the horizontal grey lines represent the mean substantive factor regression coefficient and the mean variance explained by the model with 0% of wording effect.

As in Studies 1 and 2, a simulated sample data matrix with 1000 respondents and 20 items was used to obtain a series of scatter-plots to illustrate the relationship between the uncontaminated and contaminated substantive scores across different amounts of wording effects (see Figure 4.10). The trends observed for consistent respondents (colored in black) and

inconsistent respondents (colored in red) with both models mirrored the ones found in the case of carelessness and item verification difficulty (see Studies 1 and 2). However, in this case the shift produced in the contaminated score estimates for inconsistent respondents was less pronounced with both models, and therefore the estimates were notably more accurate. This is because in this case, items responses are modified proportionally for these respondents, while in the cases of carelessness and item verification difficulty item responses are not modified proportionally, since they are changed by their corresponding inverse response option (1 = 4, 2 = 3, etc.). Regarding the correlation between the wording method factor scores and the uncontaminated factor scores, the results for the consistent respondents were similar to those found in Studies 1 and 2 (there was no correlation), but the mean method factor scores were different from zero in this case. In contrast, in the case of the inconsistent respondents the results were different than those from Studies 1 and 2, as the wording method factor scores were not correlated with the uncontaminated factor scores.

4.14.5 Structural Validity

Panel B in Figure 4.9 shows the regression coefficients associated to the substantive factor for both models and the wording factor of the RIIFA, as well as the proportion of explained variance by each model (R^2). In terms of the mean regression coefficient for the substantive factor, both models showed a similar trend to the ones displayed in Study 2, with the difference that in this case the two were highly accurate even with high amounts of acquiescence. Additionally, the regression coefficient associated to the wording factor had a mean of zero across conditions, similarly to Study 2. Further, both models tended to underestimate the proportion of variance as the amount of acquiescence increased, but the underestimation was noticeable smaller than for item verification difficulty (both models) or carelessness (1F model).

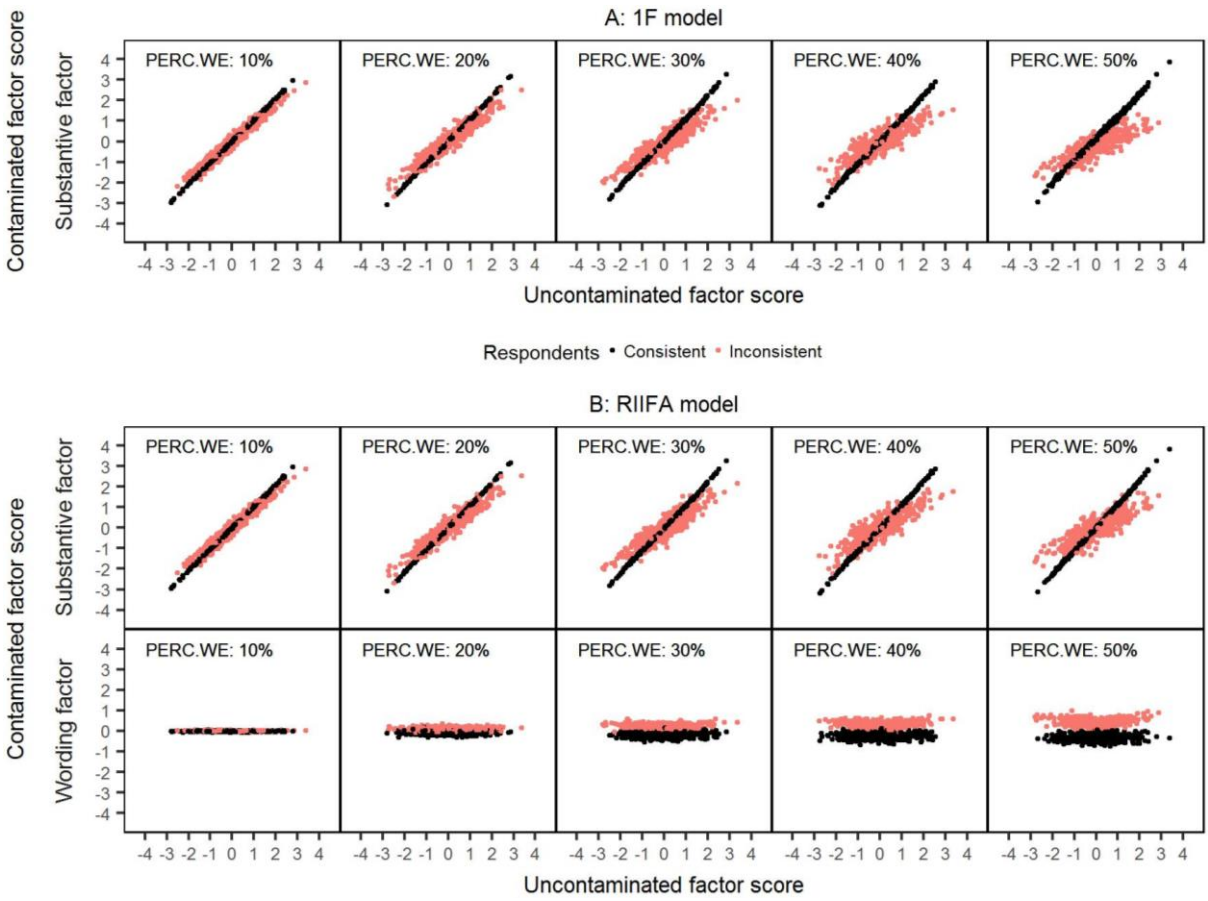

Figure 4.10. Example of recovery of the substantive factor scores with the 1F and RIIFA models in the presence of acquiescence. The data represented corresponds to simulated unidimensional data sets with 1000 respondents and 20 variables. 1F = unidimensional model with one substantive factor; RIIFA = random intercept item factor analysis model with one substantive factor and one wording method factor; PERC.WE = amount of wording effect.

4.15 Discussion

The presence of wording effects is still ubiquitous in psychological measurement. This is evidenced in the fact that researchers continue proposing and testing different strategies for controlling method effects due to inconsistent responding to polar opposite items (e.g., Plieninger & Heck, 2018; Kam, 2018; Kam & Fan, 2018). Recent research has highlighted the scarce existence of systematic studies evaluating the impact of response biases in psychometric analysis and the need to perform Monte Carlo simulation studies to shed light to this matter (Plieninger, 2016). In particular, more studies are required to evaluate whether uncontaminated

true person scores can be adequately estimated in the presence of wording effects. Moreover, despite the popularity of the RIIFA model (Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006), little is known about its behavior to estimate person scores that may be affected by wording effects. Therefore, the current study sought to fill these gaps by systematically evaluating the performance of the RIIFA model in estimating the uncontaminated person scores (and other parameters) under the influence of three wording effects: carelessness, item verification difficulty, and acquiescence (Swain et al., 2008; Weijters et al., 2013).

4.15.1 Main Findings

An initial consideration when applying the RIIFA approach concerns model *convergence*. Results suggested that the model has difficulty to disentangle wording and substantive variance if there is little information in the data set (e.g., few items) and the amount of wording effects is small. Usually, a model is of no use if the estimation does not converge (Forero et al., 2009). However, in this case it may be indicating that the impact of wording effects is minimal, and thus it is not necessary to include the random intercept in the estimated model. This is valuable information.

A fundamental step in model testing is the evaluation of *model fit*. In terms of the RMSEA and CFI values, the RIIFA model was consistently the best approach with any type of wording effect for two reasons: it was systematically superior to the 1F model across all the conditions, and always showed a good fit according to the conventional cutoff values. This is consistent with prior literature showing that the RIIFA model is superior in terms of model fit to models that only include substantive factors but not a wording factor (e.g., Abad et al., 2018; Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006; Yang et al., 2018). In fact, the fit of the RIIFA was close to the fit obtained for the 1F model with the uncontaminated datasets, indicating that it was able to properly account for the variance in the data. In contrast

but consistently with prior research (Woods, 2006), the fit of the 1F model deteriorated considerably in the presence of any type of wording effect as the amount of inconsistent responses increased in the dataset.

Regarding the recovery of the substantive *factor loadings*, in general both models showed a tendency to underestimate the factor loadings in absolute value of both PW and NW items, with any type of wording effect. In terms of the accuracy, differential trends were observed according to the type of wording effect. First, for carelessness (Study 1), both the 1F and RIIFA models showed a tendency to produce estimates biased to a greater extent for the NW items than for PW items. This was expected because we simulated carelessness specifically to NW items, as in previous research (Schmitt & Stults, 1985; Woods, 2006). That trend was accentuated with higher percentages of misresponded items. Moreover, the 1F model generally produced slightly more accurate estimates for PW items than the RIIFA, while the RIIFA model was more precise with NW items than the 1F model. All the mentioned above is valid except when examinees respond in a careless way to all the items of one type (in this case, NW items), as both the 1F and the RIIFA models will be unable to distinguish which group of items is problematic, and therefore they will produce equally biased estimates for the PW and NW items.

Second, in the presence of item verification difficulty (Study 2) and acquiescence (Study 3), both models generally produced equally accurate estimates between them and for both types of items. This can be explained because both wording effects were simulated in a balanced way: In the case of Item verification difficulty, an exact half of the inconsistent respondents misresponded to PW items and the other half to NW items. To simulate acquiescence, inconsistent respondents were randomly selected so that item responses were changed for subjects of all trait levels. As responses to PW items are mostly changed for simulees with lower trait levels, and responses to NW items are mostly changed for simulees

with higher trait levels, this produces a similar bias across both types of items. Overall, both models were more accurate with lower amounts of wording effects.

The current study focused on the recovery of the *substantive factor scores*. The results revealed that, with any type and amount of wording effect, both the 1F and the RIIFA models systematically produced accurate person score estimates for consistent respondents. This did not occur in the case of inconsistent respondents, for whom both models produced increasingly biased estimates as the amount of wording effect was greater. This differential performance across consistent and inconsistent respondents is explained because in the three studies here presented, we always simulated data matrices where the majority of the responses to the PW and NW were consistent with the 1F population model. This is what the estimated substantive factor reflects with both models. In the case of the RIIFA model, this was surprising because we expected that controlling for wording effects would lead to better person score estimates. In addition to the aforementioned results, the recovery of the substantive scores of inconsistent respondents was notably better with both models when the wording effect was Acquiescence. This was particularly noticeable in conditions with stronger wording effects.

Furthermore, the 1F and the RIIFA models performed similarly in recovering the substantive scores of inconsistent respondents when the wording effect was acquiescence or item verification difficulty. However, in the case of carelessness the 1F model was slightly superior to the RIIFA. This differential performance is related to the recovery of the factor structure because each individual item will contribute to the estimated person score proportionally to the magnitude of its substantive factor loading. In other words, when scoring a careless person who has misresponded to NW items, the 1F model will give slightly more (but not exclusive) importance to PW items (which contain correct information about the true trait levels of inconsistent examinees) than the RIIFA. In turn, the RIIFA model will give more

(but not exclusive) importance to NW items (which contain wrong information about the true trait level of inconsistent respondents) than the 1F model.

A notable finding from the three studies performed is that using the RIIFA to model wording effects produced similar results in terms of the recovery of the structural validity than "doing nothing". These results are consistent with prior research: Yang et al. (2018) applied a depression scale to a sample of Chinese adolescents, and they found that model fit improved when applying the RIIFA. However, they found that the diagnostic accuracy of the instrument was slightly better when using the raw sum scores (which would be similar to "doing nothing") than with the factor scores obtained using the RIIFA model. Maydeu-Olivares and Coffmann (2006) found similar results.

A notable finding from Studies 1 and 2 is that, in the presence of carelessness or item verification difficulty, the wording factor scores of the inconsistent respondents may reflect their uncontaminated substantive scores. This may have important implications in practice because it is common that researchers examine and interpret the correlation between a wording method factor and other measures to test validity, to identify the underlying wording effect, or even to comprehend the substantive meaning of the wording factor (e.g., Alessandri, Vecchione, Eisenberg, & Łaguna, 2015; Billiet & McClendon, 2000; DiStefano & Motl, 2006; Tomás, Oliver, Galiana, Sancho, & Lila, 2013; Ye, 2009). Therefore, we strongly recommend researchers to be especially cautious in such practice because one usually ignores the content of the wording scores.

4.15.2 Limitations and Future Research Lines

The current study has some limitations that deserve further discussion. First, although in each simulated data set only one type of wording effect was generated separately to control for other influences, in practice some of them can manifest simultaneously. That is, in an empirical sample, there may exist differences at the between-respondent level regarding the

type of wording effect influencing the response process (Grønhaug & Heide, 1992). In addition, the responses of the same examinee may be affected by different types of wording effects simultaneously (i.e., there may be different wording effects at the within-respondent level).

Another limitation of this study is that we simulated balanced scales containing the same number of PW and NW items, as has been widely recommended (e.g., Paulhus, 1991). However, prior research has shown that including different number of PW and NW items may affect parameter estimates (e.g., Plieninger, 2016). Future studies should investigate whether the RIIFA model is also robust with unbalanced scales containing fewer PW or NW items.

4.15.3 Practical Implications

The title of this article posited the question of whether modeling wording effects really help to recover uncontaminated person scores. Although findings from this study seem to point to a less than favorable answer in the case of unidimensional constructs, it is still useful to control for wording effects in these cases. First, previous studies have shown that ignoring wording effects may distort the factor structure of unidimensional constructs, leading to the emergence of separate factors for PW and NW items (e.g., Carmines & Zeller, 1979; Marsh, 1996). This can be very harmful in theory development efforts because a conceptually incorrect factor structure may be erroneously preferred over the true one (Woods, 2006). The inclusion of a random intercept factor allows us to control for this kind of detrimental effects by separating wording method variance from substantive variance. Consequently, we recommend the use of the RIIFA model always when testing a model and echo the specific recommendations made by Maydeu-Olivares and Coffman (2006) of checking the magnitude of the variance of the random intercept and of the substantive and wording factor loadings to evaluate the presence and impact of wording effects.

Another issue that may happen in practice is that the RIIFA model leads to a non-convergence solution. This can be also a valuable information because it might be indicating

that wording effects are so minimal that cannot be captured by the random intercept and that it should not be included in the model. Previous recommendations are valid for the case of multidimensional constructs with one difference: in the multidimensional scenario the distortion of the factor structure produced by wording effects may be more harmful because item may load on a wrong factor. In this case, modeling wording effects with the RIIFA will produce more accurate person scores estimates than "doing nothing" because the factor structure will be estimated adequately.

The findings from this study demonstrated that the RIIFA can successfully model the method variance generated from different types of wording effects that are not necessarily acquiescence. This is very important because researchers often erroneously interpret that wording factors measure acquiescence effects (Billiet & McClendon, 2000). The random intercept allows to model the individual use of the response scale (Maydeu-Olivares & Coffman, 2006), which might be influenced by different wording effects, including acquiescence. In this regard, we recommend that researchers be cautious when interpreting the relationships (or lack thereof) between wording factors and other measures because one might not be sure about the actual meaning or origin of these scores.

In closing, it is important to emphasize that fitting the random intercept model is not the only solution to explore wording effects. Alternative models should be tested and parameter estimates should be examined. But more importantly, we strongly echo the recommendations of other researchers that the conclusion regarding the adequacy of factor models should be based not only on statistical criteria but also on substantive and theoretical considerations (Maydeu-Olivares & Coffman, 2006; Garrido, Abad, & Ponsoda, 2016).

References

Abad, F. J., Sorrel, M. A., Garcia, L. F., & Aluja, A. (2018). Modeling general, specific, and method variance in personality measures: Results for ZKA-PQ and NEO-PI-R. *Assessment, 25*, 959-977. doi: 10.1177/1073191116667547

Alessandri, G., Vecchione, M., Eisenberg, N., & Łaguna, M. (2015). On the factor structure of the Rosenberg (1965) General Self-Esteem Scale. *Psychological Assessment, 27*, 621-635. doi: 10.1037/pas0000073

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38,* 143-156. doi: 10.1509/jmkr.38.2.143.18840

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60,* 361-370. doi: 10.1177/00131640021970592

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*, 608-628. doi: 10.1207/S15328007SEM0704_5

Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment. Beverly Hills, CA: Sage.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: a psycholinguistic processing model of verification. *Psychological review, 82*, 45-73. doi: 10.1037/h0076248

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality, 57,* 119-130. doi:10.1016/j.jrp.2015.05.004

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13,* 440-464. doi: 10.1207/s15328007sem1303_6

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods, 21,* 328-347. doi: 10.1037/met0000059

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16,* 625-641. doi: 10.1080/10705510903203573

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's minimum average partial factor retention method with categorical variables. *Educational and Psychological Measurement, 71,* 551–570. doi: 10.1177/0013164410389489

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods,* 18, 454–474. doi: 10.1037/a0030005

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods, 21,* 93. doi: 10.1037/met0000064

Gilbert, D. T. (1991), How Mental Systems Believe. *American Psychologist, 46,* 107–119. doi: 10.1037/0003-066X.46.2.107

Grønhaug, K., & Heide, M. (1992). The impact of response styles in surveys: A simulation study. *Journal of the Market Research Society, 34,* 215–230.

Hallquist, M. N. & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling*, *25*, 621-638. doi: 10.1080/10705511.2017.1402334.

Hinz, A., Michalski, D., Schwarz, R., & Herzberg, P. Y. (2007). The acquiescence effect in responding to a questionnaire. GMS Psycho-Social Medicine, 4: Doc07.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. doi: 10.1080/ 10705519909540118

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*, 99-114. doi: 10.1007/s10869-011-9231-8

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, *55*, 243–252. doi: 10.1037/h0045996

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, *39*, 103-129. doi: 10.1016/j.jrp.2004.09.009

Kam, C. C. S. (2016). Why do we still have an impoverished understanding of the item wording effect? An empirical examination. *Sociological Methods and Research*. Advance online publication. doi: 10.1177/0049124115626177

Kam, C. C. S. (2018). Novel insights into item keying/valence effect using latent difference (LD) modeling analysis. *Journal of Personality Assessment*, *100*, 389-397, doi:10.1080/00223891.2017.1369095

Kam, C. C. S., & Fan, X. (2018). Investigating response heterogeneity in the context of positively and negatively worded items by using factor mixture modeling. *Organization Research Methods*. Advance online publication. doi: 10.1177/1094428118790371

Kam, C. C. S., & Meyer, J. P. (2015a). Implications of item keying and item valence for the investigation of construct dimensionality. *Multivariate Behavioral Research*, *50*, 457–469. doi: 10.1080/00273171.2015.1022640

Kam, C. C. S., & Meyer, J. P. (2015b). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, *18*, 512–541. doi: 10.1177/1094428115571894

Kam, C., Zhou, X., Zhang, X., & Ho, M. Y. (2012). Examining the dimensionality of self-construals and individualistic–collectivistic values with random intercept item factor analysis. *Personality and Individual Differences*, *53*, 727-733. doi: 10.1016/j.paid.2012.05.023

Knowles, E. S., & Condon, C. A. (1999). Why people say" yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*, 379-386. doi: 10.1037/0022-3514.77.2.379

Marsh, H. W. (1986). The bias of negatively worded items in rating scales for young children: A cognitive-developmental phenomenon. *Developmental Psychology*, 22, 37-49.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*, 810-819. doi: 10.1037/0022-3514.70.4.810

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, 11, 344-362. doi:10.1037/1082-989X.11.4.344

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82. doi: 10.1037/1082-989X.7.1.64

Meade, A. W., & Craig, S. B. (2012). Identifying Careless Responses in Survey Data. *Psychological Methods*, *17*, 437-455 doi: 10.1037/a0028085

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171–189. doi:10.1111/j.2044-8317.1985.tb00832.x

Muthén, L. K., & Muthén, B. O. (1998-2012). Mplus user's guide (7th ed.). Los Angeles, CA: Muthén & Muthén.

Navarro-González, D., Lorenzo-Seva, U., & Vigil-Colet, A. (2016). How response bias affects the factorial structure of personality self-reports. *Psicothema*, *28*, 465-470. doi: 10.7334/psicothema2016.113

Nunnally, N., & Bernstein, I. (1994). *Psychometric Theory (3rd ed.)*. New York: McGraw-Hill.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of social psychological attitudes, Vol. 1. Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA, US: Academic Press. doi: 10.1016/B978-0-12-590241-0.50006-X

Paulhus, D. L. (2002). Socially Desirable Responding: The Evolution of a Construct. H. I., Braun, D. N. Jackson, & D. E. Wiley(Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ: Erlbaum.

Plieninger, H. (2016). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*, 32-53. doi: 10.1177/0013164416636655

Plieninger, H., & D. W. Heck (2018): A New Model for Acquiescence at the Interface of Psychometrics and Cognitive Psychology, *Multivariate Behavioral Research*. Advance online publication. doi: 10.1080/00273171.2018.1469966

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment, 25*, 1137- 1145. doi: 10.1037/a0033323

Rodebaugh, T. L., Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L., & Rapee, R. M. (2004). More information from fewer questions: the factor structure and item properties of the original and brief fear of negative evaluation scale. *Psychological Assessment, 16*, 169. doi: 10.1037/1040-3590.16.2.169

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education, 35*, 113-130. doi: 10.1080/02602930802618344

Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate behavioral research, 49*, 407-424. doi: 10.1080/00273171.2014.931800

Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology, 66*, 201–223. doi: 10.1111/j.2044-8317.2012.02049.x

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9*, 367-373. doi: 10.1177/014662168500900405

Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental

investigation. *Educational and Psychological Measurement, 51*, 67-78. doi: 10.1177/0013164491511005

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45*, 116-131. doi: 10.1509/jmkr.45.1.116

Tomás, J. M., Oliver, A., Galiana, L., Sancho, P., & Lila, M. (2013). Explaining method effects associated with negatively worded items in trait and state global and domain-specific self-esteem scales. *Structural Equation Modeling, 20*, 299-313. doi: 10.1080/10705511.2013.769394

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. (2000). *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.

Van Vaerenbergh, Y., & Thomas, T. D. (2012). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research, 25*, 195-217. doi: 10.1093/ijpor/eds021

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research, 49*, 737-747. doi: 10.1509/jmr.11.0368

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods, 18*, 320-334. doi: 10.1037/a0032121

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T.L. Leong & I. Dragos, (Eds.). *The ITC International Handbook of Testing and Assessment*. (pp. 349-363). New York: Oxford University Press.

Wetzel, E., Böhnke, J. R., & Rose, N. (2016). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement, 76*, 304-324. doi: 10.1177/0013164415591848

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*, 186. doi: 10.1007/s10862-005-9004-7

Yang, W., Xiong, G., Garrido, L. E., Zhang, J. X., Wang, M. C., & Wang, C. (2018). Factor structure and criterion validity across the full scale and ten short forms of the CES-D among Chinese adolescents. *Psychological Assessment, 30*, 1186. doi: 10.1037/pas0000559

Ye, S. (2009). Factor structure of the General Health Questionnaire (GHQ-12): The role of wording effects. *Personality and Individual Differences*, *46*, 197-201. doi: 10.1016/j.paid.2008.09.027

Zhang, X., Noor, R., & Savalei, V. (2016). Examining the effect of reverse worded items on the factor structure of the need for cognition scale. PLoS ONE 11(6): e0157795. doi:10.1371/journal.pone.0157795

Ziegler, M. (2015). "F*** you, I won't do what you told me!"–response biases as threats to psychological assessment. *European Journal of Psychological Assessment*, *31*, 153-158. doi: 10.1027/1015-5759/a000292

Chapter 5

General Discussion

Over the last decades, the consolidation of the Big Five model as the dominant paradigm in personality research (John, Naumann, & Soto, 2008) has allowed the development of a common framework of investigation and evaluation of personality traits. During this process, the evolution in the comprehension of the Big Five traits of personality, as well as the advances in measurement theory, have highlighted a series of theoretical and methodological issues that need to be addressed so that personality psychology continues to progress. This dissertation sought to address some of these issues. In the course of this process, both empirical data and Monte Carlo methods were used in the studies here presented.

Traditionally, personality traits have been measured through the administration of paper-and-pencil inventories (i.e., fixed-length tests) such as the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992). However, these personality measures present some limitations. First, they are usually very long (e.g., 240 items for the NEO-PI-R and the IPIP-NEO, respectively), and result in inefficient and time-consuming individual evaluations considering that researchers have a very limited testing time (Rammstedt & John, 2007). Second, short versions of these questionnaires are not the optimal solution because they do not preserve the psychometrics properties of the original scales (e.g., they only partially retain the original facet structure; Gignac, Bates, & Jang, 2007). Third, content and scoring keys of the

items of some of these measures are in the public domain, which is very useful for research purposes but discourages their use in evaluation contexts.

Computerized adaptive testing (CAT) has proved to be an adequate methodology for addressing the limitations of traditional testing because they improve its efficiency by only administering items tailored to the trait of the examinee (Wainer, 2000; Weiss, 1985). Pioneer attempts to measure the Big Five model adaptively have corroborated this fact in the Danish (Makransky, Mortensen, & Glas, 2013) and American (Reise & Henson, 2000) contexts. However, these studies applied unidimensional or traditional multidimensional (correlated-traits) item response theory (IRT) models to calibrate the item pools, ignoring the hierarchical structure of the Big Five personality traits that theorist and researchers assume at the present (Costa & McCrae, 1995; Soto & John, 2009). To this respect, the bifactor model has proved to be optimal in representing constructs with a hierarchical or multifaceted structure as is the case of the Big Five model (Reise, 2012). In the last decade, the development of multidimensional CATs based on the bifactor model (MCAT-B) has increased importantly. Most MCAT-Bs have been conceived to assess psychopathological measures such as depression, anxiety, and schizotypal personality (Gibbons et al., 2008, 2012, 2014; Gibbons, Weiss, Frank, & Kupfer, 2016; Moore, Calkins, Reise, Gur, & Gur, 2018; Sunderland, Batterham, Carragher, Calear, & Slade, 2017; Weiss & Gibbons, 2007; Zheng, Chang, & Chang, 2013).

In light of all the above, the two first specific goals of this dissertation were oriented towards the development of a CAT to improve the efficiency of the Big Five personality trait estimates. Specifically, the first specific goal was to develop an item pool to constitute the basis for the first CAT in the Spanish context to measure the Big Five personality traits (specifically the facets) efficiently, and to study the performance of the corresponding CATs. The second specific goal was to develop a MCAT-B to assess the Big Five personality traits adaptively and to test whether it is more adequate to this end than other traditional competing approaches: a

short scale, CAT based on a unidimensional IRT model, and a CAT based on the traditional multidimensional IRT (correlated factors) model.

Traditionally, personality research and assessment are often based on self-report measures, which are also applied to measure a number of variables in many psychological fields. This is because self-report measures present many virtues that make them the preferential framework over other approaches such as their practicality and the capability of inquire about private behaviors that would not be possible to evaluate otherwise (Paulhus & Vazire, 2005). Despite this, they have been target of criticism from the early days of psychological assessment (Allport, 1927). One of the reasons is that examinees responses might be affected by wording effects, producing inaccurate parameter estimates if wording method variance is ignored (Chan, 2009). Although an extensive body or research has examined the impact of wording effects on different psychometric properties of scales (e.g., Abad et al., 2018; Kam & Meyer, 2015a), few studies have studied their influence on the estimation of person scores (e.g., Plieninger, 2016). Even fewer studies have done it from a systematic perspective. In order to fill this gap in prior literature, the third specific goal was to examine the recovery of person score and other parameter estimates in the presence of three wording effects defined in prior literature: carelessness, item verification, and acquiescence. To do so, the random intercept item factor analysis (RIIFA) approach (Maydeu & Coffman, 2006) was used to model wording method variance.

5.1 Main Contributions of the Dissertation

5.1.1 Study 1: Calibrating a New item Pool to Adaptively Assess the Big Five

Study 1 illustrated the complete process followed to build the first Spanish CAT to measure the Big Five personality traits. In personality research, there is a debate about the measurement of facets versus domains which is known as the bandwidth-fidelity dilemma (Ones & Viswesvaran, 1996). Related to this, prior studies have shown that narrow measures

contribute to the prediction of several outcomes in different contexts of psychological evaluation (e.g., Ashton, Paunonen, & Lee, 2014). Thus, a first step to improve personality assessment through CAT was to build an item pool that considered the structure of the Big Five facets. To do so, four major phases were distinguished and correspondingly illustrated in this study.

The main core of a CAT is a wide pool of items that is calibrated (i.e., item and person parameters are known) according to an IRT model. The development and calibration of an item pool involve the performance of many psychometric analyses to guarantee the adequacy of the items during the adaptive administration (Bjorner, Chang, Thissen, & Reeve). In the pool built in this study a total of 480 items were initially designed to measure the Big Five facets. The whole pool was administered to a sample of 826 undergraduate students. The first phase was focused on obtaining evidence for content validity through the analysis of item content made by several experts in personality research and psychometrics. In the second phase, each facet was calibrated separately according to the unidimensional IRT graded response model. Item fit was also analyzed. As a result of these two steps, a pool with 360 items with good psychometric properties was obtained.

Once the item pool was properly calibrated, next steps were focused on examining the psychometric properties of the CATs. In the third phase, a post hoc simulation study was carried out to evaluate the performance of separate CATs to measure the Big Five facets. Results revealed that CATs provided highly accurate facet scores estimates, allowing to measure the personality efficiently with only 4 items per facet. This produces a significant reduction in test length, using only a third of the items in the pool. This was consistent with prior research (Reise & Henson, 2000). Finally, in the fourth phase, evidences for internal structure and convergent validity were obtained in favor of the usage of both the item pool and the separate CATs to measure the Big Five facets.

Overall, this study contributed to prior research by providing for the first time an item pool to measure the 30 facets of the Big Five model of personality in the Spanish context. As this pool presented good psychometric properties to be administered by means of an adaptive algorithm, this entails producing much more efficient individual evaluations than with the tedious paper-and-pencil tests traditionally used.

5.1.2 Study 2: Assessing the Big Five with Bifactor Computerized Adaptive Testing

Study 2 examined the feasibility of a MCAT-B to measure simultaneously the domains and facets of the Big Five model. This is a novel approach because, unlike other traditional models employed, the bifactor model is compatible with the hierarchical nature of the theoretical models of personality. Specifically, the main purpose of this study was to compare the performance of a MCAT-B with other three competing approaches traditionally used to measure the Big Five model: unidimensional CAT (UCAT), traditional multidimensional CAT (MCAT) based on the correlated-factors model, and a short scale. For this purpose, the sample and the item pool calibrated in the previous study were used. In this case, the item pool was calibrated according to the bifactor model: for each of the five broad personality traits, a bifactor model with a general factor representing the domain and several specific factors representing the corresponding facets was tested. Results revealed that all the estimated models fitted the data well. In addition, the pool continued showing good psychometric properties in terms of reliability and convergent validity.

One advantage of applying the bifactor model is that it allows to compute a series of indices based on it to study some of psychometric properties of the measures (for a detailed review on these indices, see Rodríguez, Reise, & Haviland, 2016a, 2016b). In this study, two of these indices were obtained to assess the essential unidimensionality of the domains: the explained common variance due to the effect of the general factor (ECV; Sijtsma, 2009; Ten Berge & Sočan, 2004) and the proportion of between-item correlations uncontaminated by the

specific factors (PUC; Bonifay, Reise, Scheines, & Meijer, 2015; Reise, Scheines, Widaman, & Haviland, 2013). Derived results revealed that although the influence of the specific facets was low in all the cases, the Big Five traits were essentially multidimensional constructs. The resultant item parameter distortion from fitting multidimensional (bifactor) data to a unidimensional model was high for all the domains except Extraversion, which manifested the lowest parameter bias. This suggests that the structure of this domain may be closer to unidimensionality.

Subsequently, a post-hoc simulation study was conducted to compare the performance of the four approaches mentioned above to assess the Big Five traits of personality. The four of them provided score estimates in the domains, but only the multidimensional methods (MCAT and MCAT-B) produced facet score estimates. Although in general the short scale and the UCAT performed worse than the MCAT and MCAT-B in estimating the domain scores, results for each domain were closely related to its degree of essential unidimensionality and for Extraversion the four methods were equally accurate. Regarding the estimation of the facet scores, both multidimensional methods performed similarly.

The pool usage of the three CAT based methods was also examined. Overall, for all the Big Five domains, the MCAT and MCAT-B provided a better content-balance of the pool because they selected items from all the facets in similar proportions. In contrast, the UCAT tended to represent to a great degree those facets with more highly informative items, whereas few or no items were selected for the remaining facets as occurred in the short scales. This is a limitation of these measures because it can constitute a source of model misfit (Gignac et al., 2007). These results were in line with prior research (Reise & Henson, 2000).

Finally, although both the MCAT and MCAT-B models showed a similar performance, this study highlights a number of advantages of the bifactor model that makes it a preferable approach over the traditional multidimensional model in practice. For example, with the

bifactor model one can estimate the residualized facet scores that can potentially contribute to the incremental prediction of several psychological measures above and beyond scores on the general factor (e.g., McAbee et al., 2014; Lee & Ashton, 2004).

Overall, this study contributed in two directions to the area of adaptive assessment of personality. First, the Big Five domains proved to be essentially multidimensional constructs and, therefore, the hierarchical structure of domains and facets might not be adequately evaluated with a unidimensional model. Second, and consequently, MCAT-B constitutes a preferential framework for adaptively assessing the Big Five model traits because it considers its hierarchical structure.

5.1.3 Study 3: Does modeling wording effects help recover uncontaminated person scores?

Study 3 analyzed for the first time the performance of the RIIFA approach across three types of wording effects proposed in the literature: Carelessness, item verification difficulty, and acquiescence (e.g., Swain et al., 2008). The focus of this study was the recovery of uncontaminated person scores in the context of unidimensional substantive models and categorical variables. Three additional aspects were evaluated: model fit, the recovery of factor loadings, and structural validity. To do so, Monte Carlo methods were used to conduct three studies (one per type of wording effect) with balanced scales. Three independent variables were manipulated in each of them: the amount of wording effect, sample size, and test length. The traditional model with one substantive factor (1F) was included as a baseline of "do nothing".

The results of the simulation studies showed that the models that included the RIIFA method factor were consistently able to account for the variance in the data, attaining almost perfect fit regardless of the amount of wording effects. In contrast, the 1F model showed increasingly poorer fit. These results corroborated previous findings related to the improvement of model fit when the RIIFA model is applied instead of "doing nothing" (e.g., Abad et al., 2018; Billiet & McClendon, 2000; Maydeu-Olivares-Coffman, 2006).

Furthermore, both models showed a general tendency to underestimate the magnitude (in absolute value) of the factor loadings across the three studies. In terms of the accuracy, differential trends were observed according to the type of wording effect, but in general, the differences between models were minimal (in the case or carelessness) or inexistent (with item verification difficulty and acquiescence).

Both the 1F and RIIFA models systematically produced accurate person score estimates for those respondents who answered consistently to positively worded (PW) and negatively worded (NW) items. However, it was surprising that the RIIFA model was not able to better estimate the uncontaminated person scores in comparison to the 1F model. These results were consistently found across the three types of simulated wording effects. In addition, results revealed that wording factor scores reflected the uncontaminated substantive scores of inconsistent respondents when the wording effect was carelessness or item verification difficulty. This, however, did not occurr with acquiescence. In relation to this, when the wording effect was carelessness, modeling wording effects with the RIIFA resulted in an increase of variance explained by the model over the "do nothing approach", especially with higher percentages of misresponded items. This is because, in that case, wording scores reflect to a greater deal the true trait level of inconsistent examinees with higher amounts of wording effects. However, in the case of item verification difficulty, modeling wording effects did not suppose any improvement over the "do nothing" approach. This occurs because in the case of careless respondents, the uncontaminated substantive scores are positively related to wording scores. However, in the case of item verification difficulty, for respondents with low uncontaminated substantive scores, wording scores relate negatively with the uncontaminated substantive scores, whereas for respondents with higher true substantive scores, the same relationship is positive. Consequently, both effects cancel each other out when related to a criterion variable. Contrarily, when the wording effect was acquiescence, wording scores did

not reflect the true trait level of acquiescent respondents. These results were consistent with prior research: Yang et al. (2018) applied a scale to measure depression to a sample of Chinese adolescents. When they used the RIIFA model, model fit improved in comparison to "doing nothing" to model wording effects. However, surprisingly, they found that the diagnostic accuracy of the instrument scores was slightly better when using the raw sum scores than modeling wording variance with the RIIFA approach. This may have important implications in practice because it is common that researchers examine and interpret the correlation between the wording factor scores and other substantive measures (e.g., Billiet & McClendon, 2000; DiStefano & Motl, 2006; Tomás, Oliver, Galiana, Sancho, & Lila, 2013). In light of these results, we strongly recommend researchers to be especially cautious in this practice.

5.2 Limitations and Future Research Lines

There are some limitations in this dissertation that deserve further discussion. Most of them have already been mentioned in each specific article, and thus only the most notable will be discussed here.

First, *Studies 1* and *2* overlap in the problem of the generalizability of the findings to other contexts due to the specificity of the analyzed samples. In this regard, although the intercorrelations between the Big Five personality traits were examined they were found to be consistent with previous research (Mount, Barrick, Scullen, & Rounds, 2005; van der Linden, te Nijenhuis, & Bakker, 2010), further research is required to replicate these results in other subpopulations.

Second, in *Studies 1* and *2* post hoc simulation studies were conducted and simulees' responses were drawn from the real dataset to examine the performance of the corresponding CATs. Although real data simulations are essential to evaluate how CAT procedures will operate with real respondents (Thompson & Weiss, 2011), it is necessary to carry out additional studies with live examinees to investigate their performance in real testing settings.

Third, in *Study 3*, the impact of each simulated wording effects was analyzed separately to control for other influences. However, in practice different several wording effects can manifest simultaneously (Grønhaug & Heide, 1992), that is, they may vary at the between-person level in real data sets. In addition, it may occur the same respondent manifests more than one wording effect while responding to a real questionnaire. For example, a person can be acquiescent and also have problems to verify some type of items according to his or her true trait level in a given scale.

Similarly, in *Study 3* wording effects were simulated in some specific ways that may vary in real settings. For example, carelessness was always simulated to NW items to be consistent with prior research (Schmitt & Stults, 1985; Woods, 2006), but it might also affect to PW items, or even there could be different careless respondents simultaneously, some of them presenting difficulties in answering to items of some type and the others to NW items. Future studies should handle these limitations in order to generate wording effects in more realistic fashions that can emulate true empirical settings.

5.3 Practical Implications

According to the results of this dissertation, some practical guidance can be provided to other researchers in the area of personality assessment, and more broadly, to those professionals working with self-report measures.

First, *Studies 1* and *2* provides a series of detailed guidance on how to develop an item pool and how to apply the bifactor model to measures that have a hierarchical nature, respectively. In this regard, researchers and practitioners interested in developing an item pool should follow the four phases illustrated in *Study 1* to this end. On the other hand, *Study* 2 contains a series of insights about the application of the bifactor model. One of the most important guidelines provided in that study is related to the computation of some indices based on the bifactor model that can provide valuable information in relation to the construct

evaluated. For example, in the case of *Study 2*, the computation of some indices such as the explained common variance due to the effect of the general factor lead to conclude that the Big Five traits of personality are essentially multidimensional constructs. This allowed to explain the differences in performance of different methods based on unidimensional (short scale, UCAT) and multidimensional (MCAT, MCAT-B) models.

Based on the results of *Study 3*, the use of the RIIFA model is always recommended in model testing and this dissertation echoes the specific recommendations made by Maydeu-Olivares and Coffman (2006) of checking the magnitude of the variance of the random intercept and of the substantive and wording factor loadings to evaluate the presence and impact of wording effects. In practice, the RIIFA model may produce non-converged solutions. This can be also a valuable information because it might be indicating that wording effects are minimal and cannot be captured by the random intercept. Consequently, the random intercept should not be included in the model.

As it has been shown in *Study 3*, often one will not be sure about what this wording factor is measuring and interpreting its relationship with other variables might result in detrimental consequences for theory development efforts.

In closing, it is important to emphasize that fitting the random intercept model is not the solution itself to explore wording effects. Alternative models should be tested and different parameter estimates should be examined. But more importantly, this dissertation echoes the recommendations made by other researchers about the adequacy of testing factor models based not only on statistical criteria but also substantive and theoretical considerations (Maydeu-Olivares & Coffman, 2006; Garrido, Abad, & Ponsoda, 2016).

5.4 Conclusion

Considering all the above, the current dissertation has contributed to the advancement of knowledge in the area of personality measurement, providing two main theoretical contributions:

a) It has proved the feasibility and advantages of a MCAT-B over other traditional approaches used,

b) It has brought to light an important problem of a model traditionally used to control for wording effects (the RIIFA) that arise when examinees respond inconsistently to PW and NW items. In addition, this dissertation has provided some insight about the performance of such model.

Capítulo 6

Discusión General

En las últimas décadas, la consolidación del modelo de los Cinco Grandes como paradigma dominante en la investigación de la personalidad (John, Naumann y Soto, 2008) ha permitido el desarrollo de un marco común de investigación y evaluación de los rasgos de la personalidad. Durante este proceso, la evolución en la comprensión de los cinco grandes rasgos de la personalidad, así como los avances en teoría de la medición, han puesto de relieve una serie de cuestiones teóricas y metodológicas que deben abordarse para que la psicología de la personalidad continúe progresando. Esta tesis se propuso abordar algunos de estos temas. En el curso de este proceso, se utilizaron datos empíricos y los métodos de Monte Carlo en los estudios aquí presentados.

Tradicionalmente, los rasgos de personalidad se han medido a través de la administración de inventarios en papel y lápiz (es decir, pruebas de longitud fija) como por ejemplo el Inventario de Personalidad NEO Revisado (en inglés, *NEO-PI-R*; Costa y McCrae, 1992). Sin embargo, estas medidas de personalidad presentan algunas limitaciones. Primero, generalmente son muy largos (p.e., 240 en el caso del NEO-PI-R), dando como resultado evaluaciones individuales ineficientes y que consumen mucho tiempo considerando que los investigadores disponen de un tiempo muy limitado para examinar (Rammstedt y John, 2007).

En segundo lugar, las versiones cortas de estos cuestionarios no son una solución óptima porque no conservan las propiedades psicométricas de las escalas originales (p.e., solo conservan parcialmente la estructura original a nivel de facetas; Gignac, Bates y Jang, 2007). En tercer lugar, el contenido y la forma de puntuuar los ítems de algunas de estas medidas son de dominio público, lo cual es muy útil para fines de investigación, pero desalienta su uso en contextos de evaluación.

Los test adaptativos informatizados (TAIs) han demostrado ser una metodología adecuada para abordar las limitaciones de las pruebas tradicionales porque mejoran su eficiencia al administrar únicamente los ítems adaptados a la habilidad del examinado (Wainer, 2000; Weiss, 1985). Los estudios pioneros que han tratado de medir el modelo de los Cinco Grandes de forma adaptativa han corroborado este hecho en el contexto danés (Makransky, Mortensen, y Glas, 2013) y también en el americano (Reise y Henson, 2000). Sin embargo, estos estudios aplicaron modelos de teoría de respuesta al ítem (TRI) unidimensionales o modelos multidimensionales tradicionales (de rasgos correlacionados) para calibrar los correspondientes bancos de ítems, ignorando la estructura jerárquica de los cinco grandes rasgos de personalidad que los teóricos e investigadores asumen en el presente (Costa y McCrae, 1995; Soto y John, 2009). A este respecto, el modelo bifactor ha demostrado ser óptimo para representar constructos con una estructura jerárquica o multifacética como es el caso del modelo de los Cinco Grandes (Reise, 2012). En la última década, el desarrollo de TAIs multidimensionales basados en el modelo bifactor (TAIM-B) ha aumentado de manera importante. La mayoría de los TAIM-B se han concebido para evaluar medidas psicopatológicas como la depresión, ansiedad y personalidad esquizotípica (Gibbons et al., 2008, 2012, 2014; Gibbons, Weiss, Frank y Kupfer, 2016; Moore, Calkins, Reise, Gur , y Gur, 2018; Sunderland, Batterham, Carragher, Calear, & Slade, 2017; Weiss y Gibbons, 2007; Zheng, Chang, y Chang, 2013).

A la luz de todo lo anterior, los dos primeros objetivos específicos de esta tesis se orientaron hacia el desarrollo de un TAI que para mejorar la eficiencia de las estimaciones de los Cinco Grandes rasgos de personalidad. Concretamente, el primer objetivo específico fue desarrollar un banco de ítems que consitutyese la base del primer TAI en el contexto español para medir los rasgos de personalidad de los Cinco Grandes (específicamente las facetas) de manera eficiente, y estudiar el desempeño de los TAIs correspondientes. El segundo objetivo específico fue desarrollar un TAIM-B para evaluar los rasgos de personalidad de los Cinco Grandes y evaluar si es más adecuado para este fin que otros enfoques tradicionales rivales: una escala corta, un TAI basado en un modelo de TRI unidimensional y un TAI basado en un modelo multidimensional tradicional de TRI (de factores correlacionados).

Tradicionalmente, la investigación y evaluación de la personalidad se han basado en medidas de autoinforme, los cuales son utilizados también para medir una serie de variables en distintos campos psicológicos. Esto se debe a que el autoinforme presenta numerosas virtudes que los convierten en el marco de trabajo preferente frente a otros enfoques, como su practicidad y la capacidad de indagar sobre comportamientos privados que de otra manera no sería posible evaluar (Paulhus y Vazire, 2005). A pesar de esto, han sido objeto de críticas desde los primeros días de la evaluación psicológica (Allport, 1927). Una de las razones es que las respuestas de los examinados pueden verse afectadas por los efectos de redacción, lo que da lugar a estimaciones sesgadas de de los parámetros si se ignora la varianza del método de redacción (Chan, 2009). Aunque un amplio cuerpo de investigación ha examinado el impacto de los efectos de redacción sobre diferentes propiedades psicométricas de las escalas (p.e., Abad et al., 2018; Kam y Meyer, 2015a), pocos estudios han estudiado su influencia en la estimación de las puntuaciones de las personas (p.ej., Plieninger, 2016). Aún son menos los estudios que lo han hecho desde una perspectiva sistemática. Para abordar este vacío en la literatura, el tercer objetivo específico que se propuso esta tesis fue examinar la recuperación

de la puntuación de la persona y la estimación de otros parámetros en presencia de tres efectos de redacción definidos en la literatura anterior: descuido o inatención, dificultad de verificación de los ítems y aquiescencia. Para ello, se utilizó el enfoque del análisis factorial de ítems de intercepto aleatorio (en inglés, RIIFA; Maydeu y Coffman, 2006) para modelar la varianza del método de redacción.

6.1 Contribuciones Principales de la Tesis

6.1.1 Estudio 1: Calibración de un Nuevo Banco de Ítems para Evaluar Adaptativamente los Cinco Grandes

El *Estudio 1* ilustra el proceso completo seguido para construir el primer TAI español para medir los Cinco Grandes rasgos de personalidad. En la investigación de la personalidad, existe un debate sobre la medición de facetas frente a dominios que se conoce como *bandwidth-fidelity dilema* (Ones y Viswesvaran, 1996). En relación a esto, estudios previos han demostrado que las medidas estrechas contribuyen a la predicción de varios resultados en diferentes contextos de evaluación psicológica (p.e., Ashton, Paunonen y Lee, 2014). Por lo tanto, un primer paso para mejorar la evaluación de la personalidad a través de un TAI fue la construcción de un banco de ítems que considerase la estructura de las facetas de los Cinco Grandes. Para ello, se distinguieron cuatro fases principales y se ilustraron de manera correspondiente en este estudio.

El núcleo principal de un TAI es un amplio banco de ítems que se calibran (es decir, donde se han estimado y por tanto se conocen los parámetros de los ítems y de las personas) de acuerdo con un modelo de TRI. El desarrollo y la calibración de un banco de ítems implica la realización de numerosos análisis psicométricos para garantizar la idoneidad de los ítems durante la administración adaptativa (Bjorner, Chang, Thissen, y Reeve). En el banco construido en este estudio, se diseñaron inicialmente un total de 480 ítems para medir las facetas de los Cinco Grandes. El banco completo se administró a una muestra de 826

estudiantes de grado. La primera fase se enfocó en obtener evidencia de la validez de contenido mediante el análisis del contenido de los ítems por parte de varios expertos en investigación de personalidad y psicometría. En la segunda fase, cada faceta se calibró por separado de acuerdo con el modelo de TRI unidimensional de respuesta graduada. También se analizó el ajuste del ítem. Como resultado de estos dos pasos, se obtuvo un banco con 360 ítems con buenas propiedades psicométricas.

Una vez que el banco de ítems se calibró correctamente, los siguientes pasos se centraron en examinar las propiedades psicométricas de los TAIs. En la tercera fase, se llevó a cabo un estudio de simulación post hoc para evaluar el desempeño de TAIs separados para medir las facetas de los Cinco Grandes. Los resultados revelaron que los TAIs proporcionaron estimaciones muy precisa de las puntuaciones en las facetas, lo que permite medir la personalidad de manera eficiente con solo 4 ítems por faceta. Esto produce una reducción significativa en la duración de la prueba, utilizando solo un tercio de los ítems del banco. Esto fue consistente con investigaciones previas (Reise y Henson, 2000). Finalmente, en la cuarta fase, se obtuvieron evidencias de la estructura interna y la validez convergente a favor del uso del banco de ítems y los TAIs individuales para medir las facetas de los Cinco Grandes.

En general, este estudio contribuyó a la investigación previa al proporcionar por primera vez un banco de ítems que mide las 30 facetas del modelo de personalidad de los Cinco Grandes en el contexto español. Este banco presentó buenas propiedades psicométricas para ser administrado por medio de un algoritmo adaptativo, lo que implica producir evaluaciones individuales mucho más eficientes que con las tediosas pruebas de papel y lápiz tradicionalmente utilizadas.

6.1.2 Estudio 2: Evaluación de los Cinco Grandes Mediante un Test Adaptativo Informatizado Basado en el Modelo Bifactor

El *Estudio 2* examinó la viabilidad de un TAIM-B para medir simultáneamente los dominios y las facetas del modelo de los Cinco Grandes. Este es un enfoque novedoso porque, a diferencia de otros modelos tradicionales empleados, el modelo bifactor es compatible con la naturaleza jerárquica de los modelos teóricos de la personalidad. Específicamente, el objetivo principal de este estudio fue comparar el rendimiento de un TAIM-B con otros tres enfoques competitivos tradicionalmente utilizados para medir el modelo de los Cinco Grandes: TAI unidimensional (TAIU), TAI multidimensional tradicional (TAIM) basado en el modelo de factores correlacionados, y una escala corta. Para este propósito, se utilizaron la muestra y el banco de ítems calibrados en el estudio anterior. En este caso, el banco de ítems se calibró de acuerdo con el modelo bifactor: para cada uno de los cinco grandes rasgos de personalidad, se evaluó un modelo bifactor con un factor general que representaba el dominio y varios factores específicos que representaban las facetas correspondientes. Los resultados mostraron que todos los modelos estimados se ajustan bien a los datos. Además, el banco continuó mostrando buenas propiedades psicométricas en términos de fiabilidad y validez convergente.

Una ventaja de la aplicación del modelo bifactor es que permite calcular una serie de índices basados en él para estudiar algunas de las propiedades psicométricas de las medidas (para una revisión detallada de estos índices, consultar Rodríguez, Reise y Haviland, 2016a, 2016b). En este estudio, se obtuvieron dos de estos índices para evaluar la unidimensionalidad esencial de los dominios: la varianza común explicada debido al efecto del factor general (en inglés, *ECV*; Sijtsma, 2009; Ten Berge y Sočan, 2004) y la proporción de las correlaciones entre ítems no contaminadas por los factores específicos (en inglés, *PUC*; Bonifay, Reise, Scheines, & Meijer, 2015; Reise, Scheines, Widaman, & Haviland, 2013). Los resultados derivados revelaron que aunque la influencia de las facetas específicas era baja en todos los

casos, los rasgos de los Cinco Grandes eran esencialmente construcciones multidimensionales. La distorsión provocada en los parámetros de los ítems tras ajustar datos multidimensionales (bifactor) a un modelo unidimensional fue alta para todos los dominios excepto para Extraversión, que manifestó el sesgo más bajo en relación a los parámetros estimados. Esto sugiere que la estructura de este dominio podría acercarse a la unidimensionalidad.

Posteriormente, se realizó un estudio de simulación post hoc para comparar el rendimiento de los cuatro enfoques mencionados anteriormente para evaluar los rasgos de personalidad de los Cinco Grandes. Los cuatro proporcionaron estimaciones de las puntuaciones en los dominios, pero solo los métodos multidimensionales (TAIM y TAIM-B) produjeron estimaciones de las puntuaciones en las facetas. Aunque en general la escala corta y el TAIU funcionaron peor que el TAIM y TAIM-B al estimar las puntuaciones en el dominio, los resultados individuales para cada dominio se relacionaron estrechamente con su grado de unidimensionalidad esencial, y en el caso de Extraversión, los cuatro métodos fueron igualmente precisos. Con respecto a la estimación de las puntuaciones en las facetas, ambos métodos multidimensionales se funcionaron un funcionamiento similar.

También se evaluó el uso que hacían del banco los tres métodos adaptativos. En general, para todos los dominios de los Cinco Grandes, el TAIM y TAIM-B mostraron un mejor balance del contenido del banco porque seleccionaron ítems de todas las facetas en proporciones similares. En contraste, el TAIU tendió a representar en gran medida aquellas facetas con más cantidad de ítems más informativos, mientras que se seleccionaron pocos o ningún ítem para las facetas restantes como ocurrió en las escalas cortas. Esta es una limitación de estas medidas porque puede constituir una fuente de desajuste para el modelo (Gignac et al., 2007). Estos resultados estuvieron en línea con investigaciones anteriores (Reise y Henson, 2000).

Finalmente, aunque los modelos TAIM y TAIM-B mostraron un rendimiento similar, este estudio destaca una serie de ventajas del modelo bifactor que lo convierten en un enfoque

preferible al modelo multidimensional tradicional en la práctica. Por ejemplo, con el modelo bifactor se pueden estimar las puntuaciones residuales en las facetas, las cuales pueden potencialmente contribuir a la predicción incremental de varias medidas psicológicas más allá de las puntuaciones en el factor general (p.e., McAbee et al., 2014; Lee y Ashton, 2004).

En general, este estudio contribuyó en dos direcciones al área de la evaluación adaptativa de la personalidad. Primero, los Cinco Grandes dominios demostraron ser esencialmente construcciones multidimensionales y, por tanto, podría no ser adecuado evaluar la estructura jerárquica de dominios y facetas con un modelo unidimensional. Segundo, y en consecuencia, el TAIM-B constituye un marco de referencia preferente para evaluar de manera adaptativa los rasgos del modelo de los Cinco Grandes porque tiene en cuenta su estructura jerárquica.

6.1.3 Estudio 3: Es Posible Recuperar las Puntuaciones Insesgadas de las Personas si se Modelan los Efectos de la Polaridad de los Ítems?

El *Estudio 3* analizó por primera vez el funcionamiento del modelo RIIFA en tres tipos de efectos de la polaridad de los ítems propuestos en la literatura: descuido o inatención, dificultad de verificación de los ítems y aquiescencia (p.e., Swain et al., 2008). El enfoque de este estudio fue la recuperación de las puntuaciones insesgadas de las personas en el contexto de modelos unidimensionales con un factor sustantivo y variables categóricas. Se evaluaron tres aspectos adicionales: el ajuste del modelo, la recuperación de los pesos factoriales y la validez estructural. Para ello, se utilizó el método de simulación Monte Carlo para realizar tres estudios (uno por cada tipo de efecto de la polaridad) con escalas balanceadas. Se manipularon tres variables independientes en cada uno de los estudios: la cantidad de efecto de redacción, el tamaño de la muestra y la longitud del cuestionario. El factor tradicional con un modelo sustantivo (1F) se incluyó como una línea base de "no hacer nada" para modelar estos efectos.

Los resultados de los estudios de simulación mostraron que los modelos que incluían el factor del método del RIIFA podían tener en cuenta la varianza de los datos de manera consistente, logrando un ajuste casi perfecto independientemente de la magnitud del efecto de la polaridad. En contraste, el modelo 1F mostró un ajuste cada vez más pobre. Estos resultados corroboraron los resultados anteriores relacionados con la mejora del ajuste del modelo cuando se aplica el modelo RIIFA en lugar de "no hacer nada" (p.e., Abad et al., 2018; Billiet & McClendon, 2000; Maydeu-Olivares-Coffman, 2006).

Además, ambos modelos mostraron una tendencia general a subestimar la magnitud (en valor absoluto) de las cargas factoriales en los tres estudios. En cuanto a la precisión, se observaron tendencias diferenciales según el tipo de efecto de redacción, pero en general, las diferencias entre los modelos fueron mínimas (en el caso de descuido) o inexistentes (con dificultad de verificación de los ítems y aceptación).

Tanto el modelo 1F como el modelo RIIFA produjeron sistemáticamente estimaciones precisas de la puntuación de la persona para los encuestados que respondieron de manera consistente a los ítems redactados de forma positiva y los redactados de forma negativa. Sin embargo, fue sorprendente que el modelo RIIFA no pudiera estimar mejor las puntuaciones no contaminadas en comparación con el modelo 1F. Estos resultados se encontraron constantemente en los tres estudios de simulación.

Además, los resultados revelaron que las puntuaciones del factor de redacción reflejaban las puntuaciones no contaminadas en el factor sustantivo para aquellos examinados que respondían de forma inconsistente cuando el efecto de redacción era descuido o la dificultad de verificación de los ítems. Esto, sin embargo, no ocurrió con la aquiescencia. En relación con esto, cuando el efecto de redacción fue descuido, el modelado de los efectos de redacción con el RIIFA permitió explicar una mayor cantidad de varianza al modelo frente el "enfoque de no hacer nada", especialmente cuando hubo más respuestas erróneas a los ítems.

Esto se debe a que, en ese caso, las puntuaciones de método reflejan en mayor medida el nivel de rasgo verdadero de los examinados inconsistentes con cuando los efectos de la polaridad son mayores. Sin embargo, en el caso de la dificultad de verificación de los ítems, el modelado de los efectos de redacción no supone ninguna mejora con respecto al enfoque de "no hacer nada". Esto ocurre porque en el caso de los encuestados descuidados, las puntuaciones sustantivas no contaminadas se relacionan positivamente con las puntuaciones de redacción. Sin embargo, en el caso de la dificultad de verificación del ítem, para los encuestados con puntuaciones sustantivas no contaminadas bajas, las puntuaciones de método se relacionan negativamente con las puntuaciones sustantivas no contaminadas, mientras que para los encuestados con puntuaciones sustantivas verdaderas más altas, la misma relación es positiva. En consecuencia, ambos efectos se anulan entre sí cuando se relacionan con una variable de criterio. Por el contrario, cuando el efecto de la redacción fue la aquiescencia, las puntuaciones de la redacción no reflejaron el verdadero nivel de rasgo de los encuestados que respondieron. Estos resultados fueron consistentes con investigaciones anteriores: Yang et al. (2018) aplicaron una escala para medir la depresión en una muestra de adolescentes chinos. Cuando utilizaron el modelo RIIFA, el ajuste del modelo mejoró frente a la alternativa de “no hacer nada”. Sin embargo, sorprendentemente, encontraron que la precisión diagnóstica de las puntuaciones del instrumento fue ligeramente mejor cuando se utilizaron las puntuaciones totales sumadas que cuando se modelo la varianza debida a método con el modelo RIIFA. Esto puede tener implicaciones importantes en la práctica porque es común que los investigadores examinen e interpreten la correlación entre las puntuaciones del factor método y otras medidas sustantivas (por ejemplo, Billiet y McClendon, 2000; DiStefano y Motl, 2006; Tomás, Oliver, Galiana, Sancho, y Lila, 2013). A la luz de estos resultados, recomendamos encarecidamente a los investigadores que sean especialmente cautelosos en esta práctica.

6.2 Limitaciones y Futuras Líneas de Investigación

Existen algunas limitaciones en esta tesis que merecen ser discutidas en mayor profunidad. La mayoría de ellas ya han sido mencionadas en cada artículo específico y, por lo tanto, solo las limitaciones más notables se discutirán aquí.

Primero, los *Estudios 1* y *2* comparten el problema de la generalización de los hallazgos a otros contextos debido a la especificidad de la muestra de estudio analizada. En este sentido, aunque se examinaron las intercorrelaciones entre los Cinco Grandes rasgos de personalidad y se encontró que eran consistentes con investigaciones anteriores (Mount, Barrick, Scullen y Rounds, 2005; van der Linden, te Nijenhuis, y Bakker, 2010), se requiere de investigación adicional/en el futuro para replicar estos resultados en otras subpoblaciones.

En segundo lugar, en los *Estudios 1* y *2* se realizaron estudios de simulación post hoc y por tanto se utilizaron las puntuaciones reales para examinar el rendimiento de los TAIs correspondientes. Si bien las simulaciones de datos reales son esenciales para evaluar cómo funcionarán los procedimientos TAI con los encuestados reales (Thompson y Weiss, 2011), es necesario realizar estudios adicionales con examinados en de verdad para investigar su desempeño en entornos de prueba reales.

Tercero, en el *Estudio 3*, el impacto de cada efecto de la polaridad simulado se analizó por separado para controlar otras influencias. Sin embargo, en la práctica, varios de estos efectos pueden manifestarse simultáneamente (Grønhaug y Heide, 1992), es decir, pueden variar individualmente para las distintas personas en bases de datos reales. Además, puede ocurrir que el mismo encuestado manifieste más de un efecto de redacción al responder a un cuestionario real. Por ejemplo, una persona puede ser aquiescente y también tener problemas para verificar ciertos ítems de acuerdo con su nivel de rasgo verdadero en la escala dada.

De manera similar, en el *Estudio 3*, los efectos de la polaridad se simularon de formas específicas que pueden variar en entornos reales. Por ejemplo, la falta de atención siempre se

simuló para los ítems redacatados de forma negativa para que fuera coherente con la investigación previa (Schmitt y Stults, 1985; Woods, 2006), pero también podría afectar a los ítems positivos, o incluso podría haber diferentes encuestados descuiados al mismo tiempo, algunos de ellos presentando problemas al responder a los ítems positivos y otros a los ítems negativos. Los estudios llevados a cabo en el futuro deben abordar estas limitaciones para generar efectos de redacción de manera más realista que puedan emular condiciones de evaluación realistas.

6.3 Implicaciones Prácticas

De acuerdo con los resultados de esta tesis, se puede brindar algunas orientaciones prácticas a otros investigadores en el área de la evaluación de la personalidad y, más ampliamente, a aquellos profesionales que trabajan con medidas de autoinforme.

Primero, los *Estudios 1* y *2* proporcionan una serie de guías detalladas sobre cómo desarrollar un banco de ítems y cómo aplicar el modelo bifactor a medidas que tienen una naturaleza jerárquica, respectivamente. En este sentido, los investigadores y profesionales interesados en desarrollar un banco de ítems deben seguir las cuatro fases ilustradas en el *Estudio 1* para tal fin. Por otro lado, el *Estudio 2* contiene una serie de ideas sobre la aplicación del modelo bifactor. Una de las pautas más importantes proporcionadas en ese estudio está relacionada con el cálculo de algunos índices basados en el modelo bifactor que pueden proporcionar información valiosa en relación con el constructo evaluado. Por ejemplo, en el caso del *Estudio 2*, el cálculo de algunos índices, como la varianza común explicada debido al efecto del factor general, lleva a la conclusión de que los Cinco Grandes rasgos de personalidad son esencialmente construcciones multidimensionales. Esto permitió explicar las diferencias en el rendimiento de los diferentes métodos basados en modelos unidimensionales (escala corta, TAIU) y multidimensionales (TAIM, TAIM-B).

Basado en los resultados del *Estudio 3*, se recomienda siempre el uso del modelo RIIFA cuando se evalúan distintos modelos, y esta tesis hace eco de las recomendaciones específicas hechas por Maydeu-Olivares y Coffman (2006) de verificar la magnitud de la varianza del intercepto aleatorio y de los pesos factoriales en el factor sustantivo y en el de método para evaluar la presencia y el impacto de los efectos de la polaridad. En la práctica, el modelo RIIFA puede producir soluciones no convergentes. Esto también puede ser una información valiosa porque podría indicar que los efectos de redacción son mínimos y no pueden ser capturados por el intercepto aleatorio por tanto. En consecuencia, el intercepto aleatorio no debe incluirse en el modelo. Además, se recomienda a los investigadores que sean especialmente cautelosos al interpretar las relaciones entre los factores de redacción y otras medidas sustantivas. Como se ha demostrado en el *Estudio 3*, a menudo uno no está seguro de lo que mide este factor de redacción y la interpretación de su relación con otras variables podría tener consecuencias perjudiciales para los esfuerzos orientados al desarrollo de teoría.

Para concluir, es importante enfatizar que ajustar el modelo de intercepto aleatorio no es la solución en sí misma para explorar los efectos de la polaridad. Se deben probar modelos alternativos y se deben examinar las estimaciones de distintos parámetros. Pero lo que es más importante, esta tesis se hace eco de las recomendaciones hechas por otros investigadores sobre la idoneidad de evaluar distintos modelos basándose no solo en criterios estadísticos sino también en consideraciones sustantivas y teóricas (Maydeu-Olivares y Coffman, 2006; Garrido, Abad y Ponsoda, 2016).

6.4 Conclusión

Teniendo en cuenta todo lo anterior, la tesis actual ha contribuido al avance del conocimiento en el área de la medición de la personalidad, proporcionando dos contribuciones teóricas fundamentales:

a) Se ha demostrado la viabilidad y las ventajas de un TAIM-B sobre otros enfoques tradicionales utilizados,

b) Se ha arrojado luz sobre los problemas que surgen cuando los examinados responden de manera inconsistente a los ítems positivos y negativos al utilizar un modelo tradicionalmente muy usado para controlar los efectos de polaridad de los ítems (RIIFA). Además, esta tesis ha proporcionado información sobre el funcionamiento de dicho modelo.

References

Abad, F. J., Sorrel, M. A., Garcia, L. F., & Aluja, A. (2018). Modeling general, specific, and method variance in personality measures: Results for ZKA-PQ and NEO-PI-R. *Assessment, 25*, 959-977. doi: 10.1177/1073191116667547

Abad, F. J., Sorrel, M. A., Román, F. J., & Colom, R. (2016). The relationships between WAIS-IV factor index scores and educational level: A bifactor model approach. *Psychological Assessment, 28*, 987-1000. doi: 10.1037/pas0000228

Allport, G. W. (1927). Concepts of trait and personality. *Psychological Bulletin*, *24*, 284-293.

Ashton, M. C., Paunonen, S. V., & Lee, K. (2014). On the validity of narrow and broad personality traits: A response to Salgado, Moscoso and Berges (2013). *Personality and Individual Differences, 56*, 24-28. doi:10.1016/j.paid.2013.08.019

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences, 40*, 1235-1245. doi: 10.1016/j.paid.2005.10.018

Barnette, J. J. (2000). Effects of Stem and Likert Response Option Reversals on Survey Internal Consistency: If You Feel the Need, There is a Better Alternative to Using those Negatively Worded Stems. *Educational and Psychological Measurement, 60*, 361-370, doi: 10.1177/00131640021970592

Baumgartner, H. & Steenkamp, J. B. (2001). Response style in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143–156.

Biderman, M. D., Nguyen, N. T., Cunningham, C. J., & Ghorbani, N. (2011). The Ubiquity of Common Method Variance: The Case of the Big Five. *Journal of Research in Personality, 45*, 417-29. doi:10.1016/j.jrp.2011.05.001

Chapters 2–4 include their own reference list at the end. Only citations from Chapters 1, 5, and 6 are included in this section.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*, 608-628. doi: 10.1207/S15328007SEM0704_5

Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of Life Research, 16*, 95-108. doi: 10.1007/s11136-007-9168-6

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling, 22,* 504–516. doi: 10.1080/10705511.2014.938596

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review, 82*, 45-73. doi: 10.1037/h0076248

Cauffman, E., Kimonis, E. R., Dmitrieva, J., & Monahan, K. C. (2009). A multimethod assessment of juvenile psychopathy: comparing the predictive utility of the PCL: YV, YPI, and NEO PRI. *Psychological Assessment, 21*, 528-542

Chan, D. (2009). So why ask me? Are self-report data really that bad? In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and metholodogical myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 311–338). New York: Routledge.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality, 80*, 219-251. doi:10.1111/j.1467-6494.2011.00739.x

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. Multivariate Behavioral Research, *41*, 189-225. doi: 10.1207/s15327906mbr4102_5.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. Psychological Bulletin, *136*, 1092-1122. doi: 10.1037/a0021212

Cook, K. F., Teal, C. R., Bjorner, J. B., Cella, D., Chang, C. H., Crane, P. K., ... & Raczek, A. E. (2007). IRT health outcomes data analysis project: an overview and summary. *Quality of Life Research, 16*, 121-132. doi: 10.1007/s11136-007-9177-5

Costa, P., & McCrae, R. R. (1992). *NEO PI-R manual professional.* Odessa, FL: Psychological Assessment Resources, Inc.

Costa, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment, 64*, 21-50. doi: 10.1207/s15327752jpa6401_2

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*, 475-494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and psychological measurement, 10*, 3-31. doi: 10.1177/001316445001000101

Curran, P.G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19. doi: 10. 1016/j.jesp.2015.07.006

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality, 57*, 119-130. doi: 10.1016/j.jrp.2015.05.004

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13*, 440-464. doi: 10.1207/s15328007sem1303_6

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Forbey, J. D., & Ben-Porath, Y. S. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 computerized adaptive version. *Psychological Assessment, 19*, 14-24. doi: 10.1037/1040-3590.19.1.14

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31,* 4-19. doi: 0.1177/0146621606289485

Gibbons, R. D., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2014). Development of the CAT-ANX: A computerized adaptive test for anxiety. *American Journal of Psychiatry, 171*, 187-194. doi: 10.1176/appi.ajp.2013.13020178

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology, 12*, 83-104. doi: 10.1146/annurev-clinpsy-021815-093634

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59,* 361-368. doi:10.1176/appi.ps.59.4.361.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *Archives of General Psychiatry, 69,* 1104-1112. doi:10.1001/archgenpsychiatry.2012.14

Gignac, G. E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence, 55,* 57-68. doi: 10.1016/j.intell.2016.01.006

Gignac, G. E., Bates, T. C., & Jang, K. L. (2007). Implications relevant to CFA model misfit, reliability, and the five-factor model as measured by the NEO-FFI. *Personality and Individual Differences, 43,* 1051-1062. doi:10.1016/j.paid.2007.02.024

Gignac, G.E. & Watkins, M. (2013). Bifactor Modeling and the Estimation of Model-Based Reliability in the WAIS-IV. *Multivariate Behavioral Research, 48,* 639-662. 10.1080/00273171.2013.804398.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46,* 107-119. doi: 10.1037/0003-066X.46.2.107

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe, 7,* 7-28. Tilburg, The Netherlands: Tilburg University Press.

Grønhaug, K., & Heide, M. (1992). The impact of response styles in surveys: A simulation study. *Journal of the Market Research Society, 34,* 215–230.

Haley, S. M., Ni, P., Dumas, H. M., Fragala-Pinkham, M. A., Hambleton, R. K., Montpetit, K., ... & Tucker, C. A. (2009). Measuring global physical health in children with cerebral palsy: illustration of a multidimensional bi-factor model and computerized adaptive testing. *Quality of Life Research, 18,* 359-370. doi: 10.1007/s11136-009-9447-5

Hinz, A., Michalski, D., Schwarz, R., & Herzberg, P. Y. (2007) The acquiescence effect in responding to a questionnaire. *Psychosocial Medicine, 4,* 1-9.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*, 99-114. doi: 10.1007/s10869-011-9231-8

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, *55*, 243–252. doi: 10.1037/h0045996

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. John, R. Robins, & L. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). New York, Guilford.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, *39*, 103-129. doi: 10.1016/j.jrp.2004.09.009

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51*, 78-89. doi:10.1016/j.jrp.2014.05.003

Kam, C. C. S. (2016). Why Do We Still Have an Impoverished Understanding of the Item Wording Effect? An Empirical Examination. *Sociological Methods & Research*. Advance online publication. doi:10.1177/0049124115626177

Kam, C. C. S. (2018). Novel insights into item keying/valence effect using latent difference (LD) modeling analysis. *Journal of Personality Assessment*, *100*, 389-397, doi:10.1080/00223891.2017.1369095

Kam, C. C. S., & Meyer, J. P. (2015a). Implications of item keying and item valence for the investigation of construct dimensionality. *Multivariate Behavioral Research*, *50*, 457–469. doi: 10.1080/00273171.2015.1022640

Kam, C. C. S., & Meyer, J. P. (2015b). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods, 18*, 512–541. doi: 10.1177/1094428115571894

Kam C. C. S., Zhou M. (2015). Does acquiescence affect individual items consistently? *Educational and Psychological Measurement, 75*, 764-784.

Kam, C., Zhou, X., Zhang, X., & Ho, M. Y. (2012). Examining the dimensionality of self-construals and individualistic–collectivistic values with random intercept item factor analysis. *Personality and Individual Differences, 53*, 727–733.

Kappe, R., & van der Flier, H. (2010). Using multiple and specific criteria to assess the predictive validity of the Big Five personality factors on academic performance. *Journal of Research in Personality, 44*, 142-145. doi: 10.1016/j.jrp.2009.11.002

Knowles, E. S., & Condon, C. A. (1999). Why people say" yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology, 77*, 379-386. doi: 10.1037/0022-3514.77.2.379

Koutsos, P., Wertheim, E. H., & Kornblum, J. (2008). Paths to interpersonal forgiveness: The roles of personality, disposition to forgive and contextual factors in predicting forgiveness following a specific offence. *Personality and Individual Differences, 44*, 337-348. doi: 10.1016/j.paid.2007.08.011

Lee, K., & Ashton, M. C. (2006). Further assessment of the HEXACO Personality Inventory: Two new facet scales and an observer report form. *Psychological Assessment, 18*, 182-191. doi:10.1037/1040-3590.18.2.182

Makransky, G., Mortensen, E. L., & Glas, C. A. (2013). Improving personality facet scores with multidimensional computer adaptive testing: an illustration with the NEO PI-R. *Assessment, 20*, 3-13. doi: 10.1177/1073191112437756

Maples, J. L., Guan, L., Carter, N. T., & Miller, J. D. (2014). A test of the International Personality Item Pool representation of the Revised NEO Personality Inventory and development of a 120-item IPIP-based measure of the five-factor model. *Psychological Assessment, 26*, 1070-1084. doi: 10.1037/pas0000004

Marsh, H. W. (1986). The bias of negatively worded items in rating scales for young children: A cognitive–developmental phenomenon. *Developmental Psychology*, *22*, 37–49.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifacts? *Journal of Personality and Social Psychology, 70*, 810–819.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344-362. doi:10.1037/1082-989X.11.4.344

McAbee, S. T., Oswald, F. L., & Connelly, B. S. (2014). Bifactor models of personality and college student performance: A broad versus narrow view. *European Journal of Personality*, *28*, 604-619. doi:10.1002/per.1975

McCrae, R. R., & Costa Jr., P. T. (2007). Brief versions of the NEO-PI-3. *Journal of Individual Differences, 28*, 116-128. doi: 10.1027/1614-0001.28.3.116

McCrae, R. R., & Costa, Jr, P. T. (2008). The Five-Factor theory of personality. In O. John, R. Robins & L. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 159-181). New York, Guilford

McCrae, R. R., Costa, Jr, P. T., & Martin, T. A. (2005). The NEO-PI-3: A more readable revised NEO personality inventory. *Journal of Personality Assessment*, *84*, 261-270. doi: 10.1207/s15327752jpa8403_05

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437-455. doi: 10.1037/a0028085

Moore, T. M., Calkins, M. E., Reise, S. P., Gur, R. C., & Gur, R. E. (2018). Development and public release of a computerized adaptive (CAT) version of the Schizotypal Personality

Questionnaire. *Psychiatry Research, 263,* 250–256. doi: 10.1016/j.psychres.2018.02.022

Mount, M. K., Barrick, M. R., Scullen, S. M., & Rounds, J. (2005). Higher-order dimensions of the big five personality traits and the big six vocational interest types. *Personnel Psychology, 58,* 447-478. doi: 10.1111/j.1744-6570.2005.00468.x

Navarro-González, D., Lorenzo-Seva, U., & Vigil-Colet, A. (2016). How response bias affects the factorial structure of personality self-reports. *Psicothema, 28,* 465-470. doi: 10.7334/psicothema2016.113

Nieto, M. D., Abad, F. J., Hernández-Camacho, A., Garrido, L. E., Barrada, J. R., Aguado, D., & Olea, J. (2017). Calibrating a new item pool to adaptively assess the Big Five. *Psicothema, 29,* 390–395. doi: 10.7334/psicothema2016.39

Nieto, M. D., Abad, F. J., & Olea, J. (2018). Assessing the Big Five With Bifactor Computerized Adaptive Testing. *Psychological Assessment, 30,* 1678–1690. doi: 10.1037/pas0000631

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

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior*, *17,* 609-626. doi: 10.1002/(SICI)1099-1379(199611)17:6<609:: AID-JOB1828>3.0.CO;2-K

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of social psychological attitudes, Vol. 1. Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA, US: Academic Press. doi: 10.1016/B978-0-12-590241-0.50006-X

Paulhus, D. L., & Vazire, S. (2005). The Self-Report Method. In R. W. Robins, R. Fraley & R. F. Krueger (Eds.). *Handbook of research methods in personality psychology* (pp. 224-239). New York: Guilford Press.

Plieninger, H. (2016). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement, 77,* 32-53. doi: 10.1177/0013164416636655

Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment, 25*, 1137-1145.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality, 41*, 203-212. doi:10.1016/j.jrp.2006.02.001

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667-696. doi:10.1080/00273171.2012.715555

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*, 347-364. doi: 10.1177/107319110000700404

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73,* 5-26. doi: 10.1177/0013164412449831

Revicki, D. A., Chen, W. H., & Tucker, C. (2015). Developing item banks for patient-reported health outcomes. In Reise, S. P. & Revicki, D. A. (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 334-363). New York, NY: Routledge.

Rodebaugh, T. L., Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L., & Rapee, R. M. (2004). More information from fewer questions: The factor structure and item properties of the original and Brief Fear of Negative Evaluation Scale. *Psychological Assessment, 16*, 169–181. doi:10.1037/1040-3590.16.2.169.

Rodríguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*, 223-237. doi:10.1080/00223891.2015.1089249

Rodríguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*, 137-150. doi: 10.1037/met0000045

Roszkowski, M.J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education, 35*, 113-130. doi: 10.1080/02602930802618344

Rudick, M. M., Yam, W. H., & Simms, L. J. (2013). Comparing countdown-and IRT-based approaches to computerized adaptive personality testing. *Psychological Assessment, 25*, 769-779. doi: 10.1037/a0032541

Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research, 49*, 407-424. doi: 10.1080/00273171.2014.931800

Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement, 51*, 67-78. doi: 10.1177/0013164491511005

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of personality and social psychology, 94*, 168-182. doi: 10.1037/0022-3514.94.1.168

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9*, 367-373. doi: 10.1177/014662168500900405

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120. doi: 10.1007/S11336-008-9101-0

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment, 17*, 28–43. doi: 10.1037/1040-3590.17.1.28

Soto, C., & John, P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, *43*, 84-90. doi: 10.1016/j.jrp.2008.10.002

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, *84*, 1041-1053. doi: 10.1037/0022-3514.84.5.1041

Sunderland, M., Batterham, P., Carragher, N., Calear, A., & Slade, T. (2017). Developing and validating a computerized adaptive test to measure broad and specific factors of internalizing in a community sample. *Assessment*. Advance online publication. doi: 10.1177/1073191117707817.

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45*, 116-131. doi: 10.1509/jmkr.45.1.116

Ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*, 613-625. doi: 10.1007/BF02289858

Thissen, D., Reeve, B. B., Bjorner, J. B., & Chang, C. H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research*, *16*, 109-119. doi: s11136-007-9169-5

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*, 1-9. Retrieved from http://pareonline.net/getvn.asp?v=16&n=1.

Tomás, J. M., Oliver, A., Galiana, L., Sancho, P., & Lila, M. (2013). Explaining method effects associated with negatively worded items in trait and state global and domain-specific self-esteem scales. *Structural Equation Modeling, 20*, 299-313. doi: 10.1080/10705511.2013.769394

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. Cambridge: Cambridge University Press.

Van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*, 315-327. doi: 10.1016/j.jrp.2010.03.003

Wainer, H. (2000). CATs: Whither and whence. *Psicologica, 21*, 121-133

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research, 49*, 737-747. doi: 10.1509/jmr.11.0368

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reverse item bias: An integrative model. *Psychological Methods, 18*, 320-334.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*, 774-789. doi: 10.1037/0022-006X.53.6.774

Weiss, D. J., & Gibbons, R. D. (2007). *Computerized adaptive testing with the bifactor model*. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC conference on computerized adaptive testing*. Retrieved from www .psych.umn.edu/psylabs/CATCentral/

Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment, 33*(5), 352-364. doi:10.1027/1015-5759/a000291

Wolfe, R. N., & Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational and Psychological Measurement, 55(2), 177-185. doi: 10.1177/0013164495055002002*

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*, 186. doi: 10.1007/s10862-005-9004-7

Yang, W., Xiong, G., Garrido, L. E., Zhang, J. X., Wang, M. C., & Wang, C. (2018). Factor structure and criterion validity across the full scale and ten short forms of the CES-D among Chinese adolescents. *Psychological Assessment, 30*, 1186. doi: 10.1037/pas0000559

Zheng, Y., Chang, C. H., & Chang, H. H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research, 22*, 491-499. doi:10.1007/s11136-012-0179-6

Zhang, X., Noor, R., & Savalei, V. (2016). Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PloS one, 11*(6). doi: 10.1371/journal. pone.0157795

Appendix A:
Contributed Work

This appendix contains a list of the contributed publications, presentations (oral communications and posters) in national and international congresses, and contribution to research projects until the completion of this dissertation.

A1. Main author publications

1. **Nieto, M. D.**, Abad, F. J., & Olea, J. (2018). Assessing the Big Five with bifactor computerized adaptive testing. *Psychological Assessment*, 30, 1678-1690. doi: 10.1037/pas0000631 (Q1, 5-Year Impact Factor = 4.107). (Chapter 1

2. **Nieto, M. D.**, Abad, F. J., Hernández-Camacho, A., Garrido, L. E., Barrada, J. R., Aguado, D., & Olea, J. (2017). A new item pool for the Big Five assessment: Calibration and adaptive application. *Psicothema*, 29, 390-395. doi: 10.7334/psicothema2016.391 (Q2, 5-Year Impact Factor = 1.914).

A2. Related research

1. Aguado, D., Rico, R., **Nieto, M. D.**, Xie, X. Y., Salas, E. (2019). Measuring teamwork competencies cross-culturally through computerized adaptive testing. Manuscript submitted for publication.

2. **Nieto, M. D.**, Abad, F. J., & Olea, J. (July 12-14, 2017). *Modeling response biases in computerized adaptive testing: Effects, practical implications, and illustration with the Big Five model*. Paper presented at the XV Congreso de Metodología de las Ciencias del Comportamiento y de la Salud, Barcelona, Spain.

3. **Nieto, M. D.**, Abad, F. J., & Olea, J. (July 12-14, 2017). *Assessing the Big Five using bifactor computerized adaptive testing: An empirical illustration*. Paper presented at the XV Congreso de Metodología de las Ciencias del Comportamiento y de la Salud, Barcelona, Spain.

4. Abad, F. J., **Nieto, M. D.**, & Olea, J. (July 12-14, 2017). *Medición adaptativa de la personalidad con control de deseabilidad social*. Paper presented at the XV Congreso de Metodología de las Ciencias del Comportamiento y de la Salud, Barcelona, Spain.

5. Olea, J., Abad, F. J., & **Nieto, M. D.** (July 3-7, 2017). *Medición de la personalidad mediante Tests Adaptivos Informatizados*. Paper presented at the III Congreso Nacional de Psicología, Oviedo, Spain

6. Abad, F. J., **Nieto, M. D.**, de la Fuente, J., & Olea, J. (July 3-7, 2017). *Aplicación de nuevos modelos factoriales en la medición adaptativa de la personalidad.* Paper presented at the III Congreso Nacional de Psicología, Oviedo, Spain.

7. **Nieto, M. D.**, Abad, F. J., & Olea, J. (June 14, 2017). *Evaluación adaptativa de los Cinco Grandes: aplicación y estudio de diferentes modelos psicométricos.* Poster presented at the IX Seminario de la Cátedra de Modelos y Aplicaciones Psicométricos [Captación del talento en el mundo digital], Madrid, Spain.

8. Abad, F. J., **Nieto, M. D.**, Olea, J., Hernández, A., Garrido, L. E., & Barrada, J. R. (July 27-29, 2016). *Calibration of a new item pool based on the Big Five: an application of the bi-factor model.* Paper presented at the VII European Congress of Methodology, Palma de Mallorca, Spain.

9. **Nieto, M. D.**, Abad, F. J., & Olea, J. (June 24, 2016). *Diseño y calibración de un nuevo banco de ítems basado en el modelo de los Cinco Grandes: aplicación del modelo bifactorial.* Poster presented at the VIII Seminario de la Cátedra de Modelos y Aplicaciones Psicométricos [Avances en selección de personal: Nuevas tecnologías en evaluación y medición], Madrid, Spain.

AC. **Contribution to research projects**

1. Spanish Ministry of Economy and Competitiveness project: "Multidimensional Computerized Adaptive Tests: Improving calibration and item selection algorithms" (PSI2017-85022-P)

2. Spanish Ministry of Economy and Competitiveness project: "Computerized adaptive testing based on new psychometric models" (PSI2013-44300-P).

Appendix B:
Published Version of Chapter 2

Nieto, M. D., Abad, F. J., Hernández-Camacho, A., Garrido, L. E., Barrada, J. R., Aguado, D., & Olea, J. (2017). Calibrating a new item pool to adaptively assess the Big Five. *Psicothema, 29*, 390–395. doi: 10.7334/psicothema2016.39

Psicothema

Psicothema 2017, Vol. 29, No. 3, 390-395
doi: 10.7334/psicothema2016.391

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Calibrating a new item pool to adaptively assess the Big Five

María Dolores Nieto[1], Francisco J. Abad[1], Alejandro Hernández-Camacho[1], Luis Eduardo Garrido[2], Juan Ramón Barrada[3], David Aguado[1] and Julio Olea[1]

[1] Universidad Autónoma de Madrid, [2] Universidad Iberoamericana en República Dominicana and [3] Universidad de Zaragoza

Abstract

Background: Even though the Five Factor Model (FFM) has been the dominant paradigm in personality research for the past two decades, very few studies have measured the FFM adaptively. Thus, the purpose of this research was the building of a new item pool to develop a computerized adaptive test (CAT) for personality assessment. **Method:** A pool of 480 items that measured the FFM facets was developed and applied to 826 participants. Facets were calibrated separately and item selection was performed being mindful of the preservation of unidimensionality of each facet. Then, a post-hoc simulation study was carried out to test the performance of separate CATs to measure the facets. **Results:** The final item pool was composed of 360 items with good psychometric properties. Findings reveal that a CAT administration of four items per facet (total length of 120 items) provides accurate facets scores, while maintaining the factor structure of the FFM. **Conclusions:** An item pool with good psychometric properties was obtained and a CAT simulation study demonstrated that the FFM facets could be measured with precision using a third of the items in the pool.

Keywords: Item pool, computerized adaptive testing, personality assessment, Five Factor Model, graded response model.

Resumen

Nuevo banco de ítems para evaluar adaptativamente los Cinco Grandes. **Antecedentes:** a pesar de que el Modelo de los Cinco Factores (MCF) ha sido el paradigma predominante durante las últimas dos décadas, muy pocos estudios han medido el MCF de forma adaptativa. El objetivo de esta investigación fue construir un nuevo banco de ítems para desarrollar un test adaptativo informatizado (TAI) para evaluar la personalidad. **Método:** se desarrolló un banco de 480 ítems para evaluar las facetas del MCF y se aplicó a 826 participantes. Cada faceta se calibró por separado y la selección de ítems se realizó atendiendo a que cada faceta fuese unidimensional. Después se realizó un estudio de simulación post-hoc para evaluar la eficiencia de TAIs a nivel de facetas. **Resultados:** el banco final estaba formado por 360 ítems con buenas propiedades psicométricas. Los resultados demostraron que la aplicación adaptativa de cuatro ítems por faceta proporciona puntuaciones precisas en las mismas, al mismo tiempo que se mantiene la estructura factorial del MCF. **Conclusiones:** el banco final está formado por ítems con buenas propiedades psicométricas. La aplicación adaptativa del banco permite medir la personalidad de forma eficiente a nivel de facetas utilizando una tercera parte de los ítems.

Palabras clave: banco de ítems, test adaptativo informatizado, evaluación de la personalidad, Modelo de los Cinco Factores, modelo de respuesta graduada.

Over the past 25 years the Five Factor Model (FFM) of personality traits (also called 'Big Five') has been established as the dominant paradigm in personality research, exceeding 300 publications per year (John, Naumann, & Soto, 2008). The FFM assumes a multifaceted structure with five broad personality traits (i.e., domains) each one containing several narrower traits (i.e., facets).

Although in personality research there is a debate about the measurement of facets versus domains, many studies have shown that narrow measures contribute to the prediction of several outcomes in various contexts (e.g., Ashton, Paunonen, & Lee, 2014). Thus, most personality tests developed to measure the FFM are based on facets. This is the case for the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992) and the International Personality Item Pool Representation of the NEO PI-R (IPIP-NEO; Goldberg, 1999).

Because the FFM contain many facets, these questionnaires are usually very long (e.g., 240 items for the NEO PI-R), resulting in individual assessments that are oftentimes time consuming and inefficient. As a counter measure, short versions of such scales have been proposed but these have been designed to assess the broad domains, thereby ignoring the individual facet scores and even excluding facets. For example, the NEO Five-Factor Inventory-3 (NEO-FFI-3; McCrae & Costa Jr., 2007) is a version of the NEO PI-R with 60 items taken from 28 of the 30 facet scales. Another characteristic of some personality tests like the IPIP-NEO is that the items are placed in the public domain. Although this has given rise to great advances in personality research, its use could not be recommended in evaluation contexts where examinees must not know the item content prior to the administration.

Advances in measurement with item response theory (IRT) have allowed the application of computerized adaptive testing (CAT) as an alternative to traditional tests in a variety of contexts,

Received: December 22, 2016 • Accepted: April 6, 2017
Corresponding author: Francisco José Abad
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
e-mail: fjose.abad@uam.es

including the study of personality. Pioneer attempts have been carried out recently to measure the Big Five adaptively. Two studies have performed real-data simulations using responses to the NEO-PI-R items. First, Reise & Henson (2000) found that administrating separate CATs for evaluating the FFM facets provided accurate estimates with half of the NEO PI-R items. More recently, Makransky, Mortensen, and Glas (2012) applied separate multidimensional CATs in order to measure the facets on each domain and obtained increases in the reliability of the facet scores. Also, the Tailored Adaptive Personality Assessment System (TAPAS) is a CAT used to measure the FFM in military settings in the United States (e.g., Stark, Chernyshenko, Drasgow, & White, 2012). Recently in Spain, Pedrosa, Suárez-Álvarez, García-Cueto, and Muñiz (2016) developed a CAT to assess specific personality traits of enterprising personality in young people.

The main core of a CAT is the wide pool of items that is calibrated with an IRT model (i.e., the person and item parameters are known). In the Reise and Henson (2000) and Makransky et al. (2012) studies the items of the NEO-PI-R were calibrated, thereby creating an item pool. However, because a number of phases are involved in an item pool construction, the current psychometric literature recommends other rigorous analyses that should be performed before starting the calibration such as testing the unidimensionality of the constructs and the fit at the item level (e.g., Revicki, Chen, & Tucker, 2015).

In view of all the above, we present in this study the development of an item pool to constitute the basis for the first Spanish CAT to measure the FFM facets efficiently. To do so, we identify four major steps: (a) develop items of each facet and obtain evidence for content validity, (b) calibrate each facet separately, checking the unidimensionality assumption and IRT fit, (c) test the performance of separate facet CATs, and (d) obtain evidences for internal structure and convergent validity. Thus, the specific purposes of this study were (a) to design, calibrate, and validate a new item pool based on the FFM and (b) to study the performance of CATs to measure the FFM facets more efficiently.

Method

Participants

A sample of 871 psychology undergraduate students participated voluntarily in the study. The sampling was intentional. Preliminary analyses revealed that a low percentage of the participants (45 respondents, 5.16% of the initial sample) presented careless, invalid or atypical responses according to multiple criteria described in the data analysis section and were consequently excluded. The final sample was composed of 826 individuals aged 17 to 50 years (M = 20.06, SD = 3.73), of which 696 were female (70.91%). For some analyses, the whole sample was randomly divided into two datasets with equal size (n = 413), one for applying exploratory statistical analysis (model-derivation sample) and the other one for validating statistical results (validation sample). The University Research Ethics Committee granted approval for the present study. The full anonymized data set is available from the authors upon request.

Instruments

Personality item pool. According to the traditional descriptions of the FFM facets, four independent experts in personality assessment and psychometrics developed an initial pool of 480 items (16 per facet) in Spanish language. The recommendations for item pool building were followed (e.g., Revicki et al., 2015). Then, each expert reviewed the item content of the whole pool and redundant statements were excluded and replaced by new ones. The statements were administered using a five-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). A Spanish philologist revised the items and corrected grammar, spelling and style errors. Table 1 shows facets 1 to 6 for each domain.

Directed questions scale. A scale of 12 Likert-type items (1 = strongly disagree, 5 = strongly agree) directing participants to give specific responses (e.g., "If you are reading this question, please mark 'Disagree'") was applied to measure inattention. Scale scores were obtained by summing the correct responses.

NEO-FFI-3. The NEO-FFI-3 inventory, a 60-item version of the NEO-PI-3 (McCrae, Costa Jr, & Martin, 2005) to measure the FFM domains, was included to obtain evidences for convergent validity of the new item pool. The NEO-PI-3 is a revision of the NEO PI-R. Due to there are no Spanish versions of the NEO-PI-3 and the NEO-FFI-3 questionnaires, 59 of the 60 items of the NEO-FFI-3 were selected from the Spanish version of the NEO-PI-R (Cordero, Pamos, & Seisdedos, 2008). The remaining item was translated from the English version of the NEO-FFI-3.

Procedure

The items from the personality item pool, the Directed questions scale and the NEO-FFI-3 were used to create two booklets that were administered in two sessions in a counterbalanced order. Participants completed the items within an official system of data collection in a faculty of Psychology whose purpose is the participation of students in research projects in exchange for academic compensation.

Table 1
Five Factor model: Domains and facets

	Domain				
Facet	**Neuroticism**	**Extraversion**	**Openness**	**Agreeableness**	**Conscientiousness**
1	Anxiety	Warmth	Fantasy	Trust	Competence
2	Angry/hostility	Gregariousness	Aesthetics	Straightforwardness	Order
3	Depression	Assertiveness	Feelings	Altruism	Dutifulness
4	Self-consciousness	Activity	Actions	Compliance	Achievement striving
5	Impulsiveness	Excitement seeking	Ideas	Modesty	Self-discipline
6	Vulnerability	Positive emotions	Values	Tender-mindedness	Deliberation

Data analysis

Evidence for content validity. Evidence for content validity of the personality item pool was obtained. Thirty-six experts in personality research and psychometrics were asked to select the facet to which each item belonged. Each expert evaluated the items from two domains. The level of congruence between the experts for each item was measured as the percentage of classification agreement for its most chosen facet. After excluding the responses from experts with low reliability (i.e., percentage of congruence lower than 70% in at least one domain), items with less than 50% of classification in their corresponding theoretical facet were removed from the pool.

Personality item pool IRT calibration. Psychometric properties of the pool were analyzed by fitting the unidimensional graded IRT response model (Samejima, 1969) to each subset of items measuring the same facet. First, some indexes were examined in order to screen out data for careless, invalid or atypical responses (i.e., score below 9 points on the Directed questions scale, double responses in more than three items, more than 10 missing values on the personality items, outliers regarding the number of consecutive identical responses).

For each facet, the unidimensionality assumption was tested on the model-derivation sample by applying parallel analysis (PA) and the unidimensional factor model with the polychoric correlation matrix and the robust unweighted least squares (ULSMV) estimator. If unidimensionality was not tenable according to PA or some variables had very low factor loadings, items were iteratively removed until the unidimensionality assumption was met and all the items had factor loadings larger than .2. For purposes of achieving unidimensionality, the highest residual correlation was identified and the item with the smaller loading in this pair was deleted. At the end of the iterative process, PA and the comparative fit index (CFI) were used, as recommended in Garrido, Abad, & Ponsoda (2016) to assess the unidimensionality of facets in the cross-validation sample. The conventional cutoff values for the CFI, are .90 or greater for acceptable fit, and .95 or greater for good fit (Hu & Bentler, 1999).

The selected subset of unidimensional items of each facet was calibrated separately according to the graded IRT response model using the Metropolis-Hastings Robbins-Monro algorithm (MHRM; Cai, 2010a, 2010b) on the whole sample. Item fit was tested on the sample with complete response patterns using the polytomous variant of the S-χ^2 index (Orlando & Thissen, 2000) with the Benjamini-Hochberg adjustment to control Type I error (Benjamini & Hochberg, 1995). Finally, the IRT maximum a posteriori (MAP; Embretson & Reise, 2000) pool facet scores and the standard errors (*SE*s), indicating the precision of trait estimates (θ), were obtained for each individual in each facet. IRT marginal reliabilities for pool facet scores were also obtained (Brown & Croudace, 2015; p. 314).

Performance of the CAT. A post hoc simulation study was carried out to analyze the performance of the CATs in measuring the FFM facets. We simulated a separate CAT for each facet using the item responses obtained from the respondents. Since omissions are not allowed in CATs, the response vectors were completed using item and respondent estimated parameters obtained in the previous calibration step. The CAT algorithm started by selecting the item that maximized the Fisher information at $\theta = 0$ for all the respondents. Then, attending to a respondent answer, the MAP θ estimate was obtained. The next item selected was the one that maximized the Fisher information evaluated at the θ estimate. These steps were repeated until the algorithm stopped when four items were administered. Then, the final CAT facet score was estimated using the MAP method.

Different criteria were used to analyze the precision of the CATs. For each facet, the correlation between the CAT and the pool scores were obtained. We also obtained the empirical reliability and the median of the *SE* across examinees for each CAT score.

Evidence for internal structure and convergent validity for pool and CAT facet scores. First, evidence based on the factorial structure of the pool facet scores was obtained. PA with Pearson correlations was used to verify that the suggested number of factors was five as expected (one factor per personality domain). Next, we applied exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009) with the maximum likelihood estimator. Unlike exploratory factor analysis, ESEM models can include both exploratory and confirmatory methods (e.g., correlated error terms). Using the model-derivation sample, we defined five correlated ESEM factors corresponding to the five domains. The Oblimin rotation method was used. Since modification indices suggested some correlated residuals, a new model including them was tested using the cross-validation dataset. Again, PA and the CFI were used for model evaluation. Additionally, the same ESEM factor model was used to test the internal structure of CAT facet scores. Factor congruence coefficients were obtained to study the similarities of the factorial structure obtained with pool and CAT scores.

Following the previous step, pool and CAT domain scores were obtained as an average of the correspondent six facet scores. Composite reliabilities for domain scores were estimated from the ESEM models as the squared correlation between the domain trait score and the corresponding latent factor (Raykov, 1997). Finally, evidence for convergent validity was obtained by computing the correlations between the CAT and the pool domain scores with the NEO-FFI-3 raw scores.

All the analyses were performed with M*plus* 7 (Muthén & Muthén, 1998-2012) and the R packages *psych* (Revelle, 2016), *mirt* (Chalmers, 2012), and *mirtCAT* (Chalmers, 2016).

Results

Evidence for content validity. Two experts out of 36 were excluded by their low percentage of congruence (below 70%) in the Extraversion domain. After excluding these experts, the average percentages of congruence by domain were 84% for Neuroticism, 86% for Extraversion, 93% for Openness, 89% for Agreeableness, and 86% for Conscientiousness. Twenty-five items out of 480 were removed from the item pool by their low percentage of classification in the theoretical facet (less than the 50%). After excluding these items, the average percentages of classification accuracy by domain were 89% for Neuroticism, 87% for Extraversion, 94% for Openness, 90% for Agreeableness, and 89% for Conscientiousness.

Personality item pool IRT calibration. Out of 871 participants 45 were excluded from the sample of analysis because they presented careless, invalid or atypical responses. Missing data rate for item nonresponse was very low with a maximum value of 2%.

Out of 455 items 95 were removed in order to preserve the unidimensionality of each facet. The largest number of excluded items in one facet was 7 (i.e., in the Assertiveness, Straightforwardness, and Dutifulness facets). For the retained items, the unidimensionality assumption was always tenable according to PA. The unidimensional solution showed acceptable fit according to the CFI, which was equal or above .90 in 80% of the cases and equal to or higher than .85 in the remaining facets (except for Tender-mindedness, CFI = .62). PA indicated that the 67% of the facets were unidimensional. In the remaining facets, PA suggested a two-factor solution (except for Excitement seeking that PA indicated three factors). In this cases, the scree test revealed that the second empirical eigenvalue was barely greater than the random eigenvalue. All the item factor loadings on the unidimensional solutions were statistically significant ($p<.05$), with average loadings ranging from .45 to .73.

Within the framework of the IRT, only 4 items out of 360 were identified as misfitting to the graded response model according to the S-χ^2 index. The *a-parameter* of the items showed adequate positive values ranging from 0.35 to 3.86 ($a^- = 1.51$), with 23% of them being highly discriminative (i.e., $a > 2$).

Figure 1 illustrates the information and *SE* for each θ pool facet scores. For θ between –3 and 3, the *SE*s for almost all the facets, except Compliance and Dutifulness, were lower than .5, which is approximately equivalent to a reliability coefficient of .75. This indicates that the items provide good information across the different traits levels of each facet, except for the two facets mentioned. Regarding marginal reliability, all facet scores presented values equal to or above .72. Average reliabilities for pool facet scores within a domain were .89, .90, .88, .85 and .86 for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness, respectively.

Performance of the CAT. Correlations between each CAT facet scores and pool facet scores were high for all the facets with values ranging from .92 to .98 (r = .95). For most facets, the median of the participants' *SE* was lower than .4. Only Ideas Mdn_{SE}=.41), Compliance (Mdn_{SE}= .48), Tender-mindedness (Mdn_{SE}= .41), and Dutifulness (Mdn_{SE}= .53) presented higher values. Regarding marginal reliability, most facet scores presented values equal or above .7, except the Dutifulness facet with a value of .68. Average reliabilities for pool facet scores within a domain were .82, .86, .81, .79 and .79 for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness, respectively.

Evidence for internal structure and convergent validity for pool and CAT facet scores. As expected, PA based on the analysis of the pool facet scores suggested five factors. Thus, a five-factor

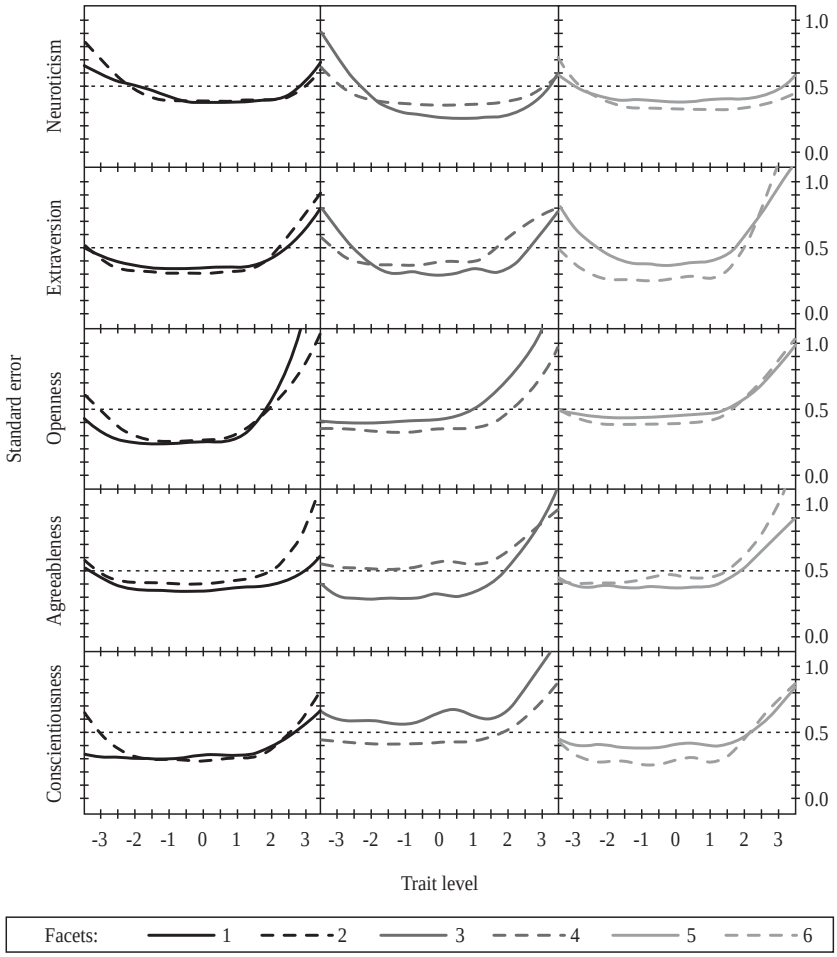

Figure 1. Standard error (*SE*) across the trait level for the facets of each domain of the FFM. *SE* equal to .50 is indicated with a dotted line. The facets 1 to 6 of each domain are specified in Table 1

exploratory model was first specified for the ESEM analyses in the model-derivation sample. This model was then modified adding six correlated residuals according to modification indexes above 40. Correlated residuals were theoretically meaningful (e.g., a negative correlation between Deliberation and Impulsiveness) and were replicated in the validation sample in which the modified model fit was acceptable: CFI was .91 and PA indicated a five-factor solution.

In the final modified model, almost all the facet scores loaded higher and significantly on its respective domain factor. These loadings were medium-high sized with values above .40 (M = .61). Only the Social anxiety and Deliberation facets presented values below .40 (.35, and .31, respectively). Regarding cross-loadings, most of them were on the Extraversion (Depression: –.33, Social anxiety: –.63, Impulsiveness: .45, Actions: .39, Trust: .35, and Deliberation: –.43), Agreeableness (Angry/hostility: –.35, Feelings: .37, Dutifulness: .35, and Deliberation: .30), and Openness (Emotions seeking: .38; Order: –.36) domains. Also Activity and Competence facets cross-loaded .33 and –.44 on Conscientiousness and Neuroticism, respectively. Average cross-loading (in absolute value) was low (.14).

The factor correlation matrix showed that Neuroticism correlated negatively with Extraversion (r = –.28; p<.001), and Conscientiousness (r = –.21; p<.001). Additionally, Extraversion also correlated, positively, with Openness (r = .24; p<.001) and Conscientiousness (r = .23; p<.001). Conscientiousness was also correlated with Openness (r = .12; p<.001) and Agreeableness (r = .10; p<.001). The remaining correlations were small ($|r|$ < .06).

When the ESEM was applied to the CAT facet scores, the results were highly similar (i.e., congruence coefficients were .99 for each of the five factors). Composite reliabilities for pool domain scores were acceptable and ranged from .75 (Agreeableness) to .87 (Extraversion). Reliabilities for CAT domain scores were inferior as expected but acceptable and ranged from .70 (Openness) to .86 (Extraversion). According to the Spearman-Brown formula and the pool composite reliabilities, it must be noted that in order to obtain these 24-item length CAT domain score reliabilities, 56 items would be required, in average, in a fixed form.

Finally, correlations between the pool domain scale scores and the NEO-FFI-3 raw scores were good. The Extraversion and Neuroticism domains presented the highest convergent validity values (r = .88 and .86, respectively). In the case of Openness and Agreeableness scales the value was similar (r =.83), and Conscientiousness presented the lowest value (r =.80). Convergent validity for the CAT domain scale scores with the NEO-FFI-3 were only slightly inferior (the largest difference, .02, was for Neuroticism).

Discussion

Recent studies in personality have investigated the possibility of obtaining accurate personality facet scores with CATs (e.g., Makransky et al., 2012). The purpose of this research was to build a new personality item pool and develop the first Spanish CAT based on the FFM facets. Analyses were performed at the facet-level. This is one of the key aspects of this study because recent research has shown that facet-level analysis increases the predictive validity of personality scores (Ashton et al., 2014).

In this study a pool of items for personality assessment is provided and efficiently administered with CAT. Although there are several commercial paper-and-pencil tests for assessing the FFM, this might be an important contribution to the evaluation of personality in applied settings where short-time assessments are required and the item content should be unknown to the examinees prior to administration.

Four main steps are distinguished in this study. First, item statements were developed and evidence for content validity was obtained via the evaluation of experts. Second, each facet was calibrated separately according to the Samejima graded response model. Unidimensionality of facets was guaranteed through a strict iterative analysis procedure and almost all the items showed adequate fit to the Samejima graded response model. In terms of precision, the facet scales showed generally good reliability with small SE over a wide range of θ. In line with previous studies (e.g., Benet-Martínez & John, 1998) and the NEO PI-R manuals, the facets of the Neuroticism, Extraversion and Openness domains were, on average, the most reliable.

Third, a CAT simulation study revealed that using separate 4-item CATs to assess the facets (i.e., with an administration of 120 items), facet scores are estimated accurately with low SEs in most cases. Finally, internal structures of the pool and the CAT were analyzed obtaining similar results: facets in both instruments measured the narrow traits of their corresponding FFM domains. Some facets loaded on more than one domain (e.g., Angry/hostility was designed to measure a subdomain of Neuroticism and was also an indicator of Agreeableness). This is consistent with previous studies that have shown that an important part of the variance of the facets scales is due to different domains (e.g., Abad, Sorrel, García, & Aluja, in press). In addition, both the item pool and CAT scores showed good convergent validity with the NEO-FFI-3 questionnaire.

One limitation of the current study is the generalizability of the results to other samples, although the intercorrelations found between the five personality factors are consistent with previous research. For example, Neuroticism correlated negatively with Extraversion and Conscientiousness, and Extraversion also correlated positively with Openness (e.g. Mount, Barrick, Scullen, & Rounds, 2005; Van der Linden, te Nijenhuis, & Bakker, 2010). Furthermore, domains such as Neuroticism and Openness showed lower correlations. However, due to the fact that the sample consisted of psychology undergraduate students, we are aware that the results may not be generalized to other sub-populations (e.g., clinical, workforce).

Recent research has suggested that multidimensional IRT models and multidimensional CATs may increase the precision of personality trait scores (e.g., Makransky et al., 2012). In this regard, future research with the presently developed item pool should be oriented toward the application of multidimensional models in the calibration and adaptive administration phases.

Acknowledgements

The research has been funded by the Ministry of Economy and Competitivity of Spain (PSI2013-44300-P), and the UAM-IIC Chair «*Psychometric Models and Applications*».

References

Abad, F. J., Sorrel, M. A., García, L. F., & Aluja, A. (in press). Modeling general, specific, and method variance in personality measures. Results for ZKA-PQ and NEO-PI-R. *Assessment*. doi: 10.1177/1073191116667547

Ashton, M. C., Paunonen, S. V., & Lee, K. (2014). On the validity of narrow and broad personality traits: A response to Salgado, Moscoso, and Berges (2013). *Personality and Individual Differences, 56*, 24-28. doi: 10.1016/j.paid.2013.08.019

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397-438. doi: 10.1080/10705510903008204

Benet-Martínez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75*, 729-750. doi: 10.1037/0022-3514.75.3.729

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*, 289-300. doi: 10.2307/2346101

Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT. In Reise, S. P. & Revicki, D. A. (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (pp. 334-363). New York, NY: Routledge.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hasings Robbins-Monro algorithm. *Psychometrika, 75*, 33-57. doi: 10.1007/s11336-009-9136-x

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*, 307-335. doi: 10.3102/1076998609353115

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1-29. doi: 10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*, 1-39. doi: 10.18637/jss.v071.i05

Cordero, A., Pamos, A., & Seisdedos, N. (2008). NEO PI-R, Inventario de Personalidad NEO Revisado [Revised NEO Personality Inventory]. *Madrid: TEA Ediciones*.

Costa, P., & McCrae, R. R. (1992). *NEO PI-R manual professional*. Odessa, FL: Psychological Assessment Resources, Inc.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, *21*, 93. doi: 10.1037/met0000064

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt & F. Ostendorf (Eds.), *Personality Psychology in Europe, 7*, 7-28. Tilburg, The Netherlands: Tilburg University Press.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi: 10.1080/10705519909540118

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. John, R. Robins & L. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, Guilford.

Makransky, G., Mortensen, E. L., & Glas, C. A. (2012). Improving personality facet scores with multidimensional computer adaptive testing: An illustration with the NEO PI-R. *Assessment, 20*, 3-13. doi: 10.1177/1073191112437756

McCrae, R. R., Costa, Jr, P. T., & Martin, T. A. (2005). The NEO-PI-3: A more readable revised NEO personality inventory. *Journal of Personality Assessment*, *84*, 261-270. doi: 10.1207/s15327752jpa8403_05

McCrae, R. R., & Costa Jr., P. T. (2007). Brief versions of the NEO-PI-3. *Journal of Individual Differences, 28*, 116-128. doi: 10.1027/1614-0001.28.3.116

Mount, M. K., Barrick, M. R., Scullen, S. M., & Rounds, J. (2005). Higher-order dimensions of the big five personality traits and the big six vocational interest types. *Personnel Psychology*, *58*, 447-478. doi: 10.1111/j.1744-6570.2005.00468.x

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide (7th ed.)*. Los Angeles, CA: Muthén & Muthén.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50-64. doi: 10.1177/01466216000241003

Pedrosa, I., Suárez-Álvarez, J., García-Cueto, E., & Muñiz, J. (2016). A computerized adaptive test for enterprising personality assessment in youth. *Psicothema, 28*, 471-478. doi: 10.7334/psicothema2016.68

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173-184. doi: 10.1177/01466216970212006

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*, 347-364. doi: 10.1177/107319110000700404

Revelle, W. (2016). Procedures for personality and psychological research. Evanston, Il: Northwestern University.

Revicki, D. A., Chen, W. H., & Tucker, C. (2015). Developing item banks for patient-reported health outcomes. In Reise, S. P. & Revicki, D. A. (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 334-363). New York, NY: Routledge.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores. Psychometrika*, *34 (Suppl. 1)*, 1-97. doi: 10.1007/BF02290599

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods*, *15*, 463-487. doi: 10.1177/1094428112444611

Van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*, 315-327. doi: 10.1016/j.jrp.2010.03.003

Appendix C:
Published Version of Chapter 3

Nieto, M. D., Abad, F. J., & Olea, J. (2018). Assessing the Big Five With Bifactor Computerized Adaptive Testing. *Psychological Assessment, 30,* 1678–1690. doi: 10.1037/pas000063

American Psychological Association

Psychological Assessment

1040-3590/18/$12.00

2018, Vol. 30, No. 12, 1678–1690
http://dx.doi.org/10.1037/pas0000631

Assessing the Big Five With Bifactor Computerized Adaptive Testing

María Dolores Nieto, Francisco J. Abad, and Julio Olea
Autonomous University of Madrid

Multidimensional computerized adaptive testing based on the bifactor model (MCAT-B) can provide efficient assessments of multifaceted constructs. In this study, MCAT-B was compared with a short fixed-length scale and computerized adaptive testing based on unidimensional (UCAT) and multidimensional (correlated-factors) models (MCAT) to measure the Big Five model of personality. The sample comprised 826 respondents who completed a pool with 360 personality items measuring the Big Five domains and facets. The dimensionality of the Big Five domains was also tested. With only 12 items per domain, the MCAT and MCAT-B procedures were more efficient to assess highly multidimensional constructs (e.g., Agreeableness), whereas no differences were found with UCAT and the short scale with traits that were essentially unidimensional (e.g., Extraversion). Furthermore, the study showed that MCAT and MCAT-B provide better content-balance of the pool because, for each Big Five domain, items from all the facets are selected in similar proportions.

Public Significance Statement
The present study illustrates the calibration procedure of an item pool to measure the Big Five personality traits according to the bifactor model. In addition, it is suggested that a multidimensional computerized adaptive test based on the bifactor model is more advantageous to assess the Big Five than other competing approaches (unidimensional computerized adaptive test, a multidimensional computerized adaptive test based on the correlated-factors model, and a short scale).

Keywords: personality assessment, Big Five, item response theory, multidimensional computerized adaptive testing (MCAT), bifactor model

The Big Five model of personality traits has been established as the dominant paradigm in personality research, exceeding 300 publications per year (John, Naumann, & Soto, 2008). The Big Five model assumes a hierarchical multifaceted structure with five broad personality traits (i.e., domains) each one containing six narrower traits (i.e., facets). Although in personality research there is a debate about the measurement of facets versus domains (Salgado et al., 2015), many studies have shown that narrow measures contribute to the prediction of several outcomes in various contexts (e.g., Ashton, Paunonen, & Lee, 2014; McAbee, Oswald, & Connelly, 2014; O'Connor & Paunonen, 2007). Thus, major personality inventories based on the 30 Big Five facets are usually very long because they contain many items to assess each facet. This is the case for the NEO Personality Inventory-3 (NEO-PI-3; McCrae, Costa, & Martin, 2005) with a total of 240 items (i.e., eight per facet) and the International Personality Item Pool Representation of the NEO-PI-R (IPIP-NEO; Goldberg, 1999) with 300 items (i.e., 10 per facet). Consequently, these questionnaires lead to individual assessments that are inefficient and time-consuming and are not recommended in short-time applications or evaluation contexts where various questionnaires need to be applied. As a countermeasure, short versions of such scales have been developed. For example, the NEO Five-Factor Inventory-3 (NEO-FFI-3; McCrae & Costa, 2007) is a 60-item version of the NEO-PI-3 (McCrae et al., 2005), although there are others. Likewise, brief versions of the IPIP-NEO have been developed, such as the IPIP-NEO-120 (e.g., Johnson, 2014; Maples, Guan, Carter, & Miller, 2014). However, these shortened questionnaires have been designed to assess the broad domains, thereby ignoring the individual facet scores and even excluding some facets. Consequently, they are less accurate than the original versions, have less convergent validity with their parent scales as the number of items decreases, and only partially retain the original facet structure (Gignac, Bates, & Jang, 2007; Johnson, 2014; McCrae & Costa, 2007).

Advances in measurement with item response theory (IRT) have allowed the application of computerized adaptive testing (CAT), improving the efficiency of traditional testing by only administering items tailored to the ability of the examinee. In personality research, pioneer attempts have been conducted to measure the Big

This article was published Online First August 30, 2018.

María Dolores Nieto, Francisco J. Abad, and Julio Olea, Department of Social Psychology and Methodology, Autonomous University of Madrid.

This research was partially supported jointly by Grants PSI2013-44300-P and PSI2017-85022-P from Ministerio de Economía y Competitividad (Spain) and the UAM-IIC Chair « Psychometric Models and Applications ».

Correspondence concerning this article should be addressed to María Dolores Nieto, Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, C/ Iván Pavlov, 6, Madrid 28049, Spain. E-mail: mariadolores.nieto@inv.uam.es

Five adaptively using CAT based on unidimensional (UCAT; Nieto et al., 2017; Reise & Henson, 2000) and multidimensional (correlated-factor) models (MCAT; Makransky, Mortensen, & Glas, 2013). These studies have shown high gains in efficiency over the administration of the complete test. On another hand, the interest in the bifactor model has increased dramatically because of its effectiveness to represent multifaceted constructs such as the Big Five personality traits (Reise, 2012). Indeed, Abad, Sorrel, García, and Aluja (2016) have endorsed the potential of MCAT based on the bifactor model (MCAT-B) for this purpose. However, the bifactor model has not been applied so far to adaptively assess the Big Five.

In this study, we propose that applying MCAT-B can provide efficient estimates of the Big Five domains and facets. In addition, we suggest that MCAT-B can provide more accurate estimates than other approaches (e.g., short scales, UCAT, and MCAT). The article is structured as follows: First, we will outline some issues about the evaluation of personality with CAT. We then provide a short background about recent applications of MCAT-B. Next, we will describe the procedure followed in this study to calibrate items according to the bifactor model in order to later apply MCAT-B. Then, we will evaluate the efficiency of score estimates on the Big Five using four different procedures for each domain (a short scale, UCAT, MCAT, and MCAT-B). Finally, we will address practical implications of adaptively assessing the Big Five personality traits. The analyses proposed in this study will be carried out using a new item pool designed to evaluate the Big Five model.

Assessing Personality With Computerized Adaptive Testing

The application of CAT to measure personality has increased over the last decades (e.g., Forbey, & Ben-Porath, 2007; Rudick, Yam, & Simms, 2013; Simms, & Clark, 2005). Specifically, in the case of the Big Five model, CAT developments have been based on the unidimensional IRT (UIRT) model to assess a single facet at a time (see Figure 1, Model A). The UIRT model assumes that there is a single primary latent dimension that explains the correlations between items. In this regard, Reise and Henson (2000) found that evaluating the facets of the NEO PI-R separately with

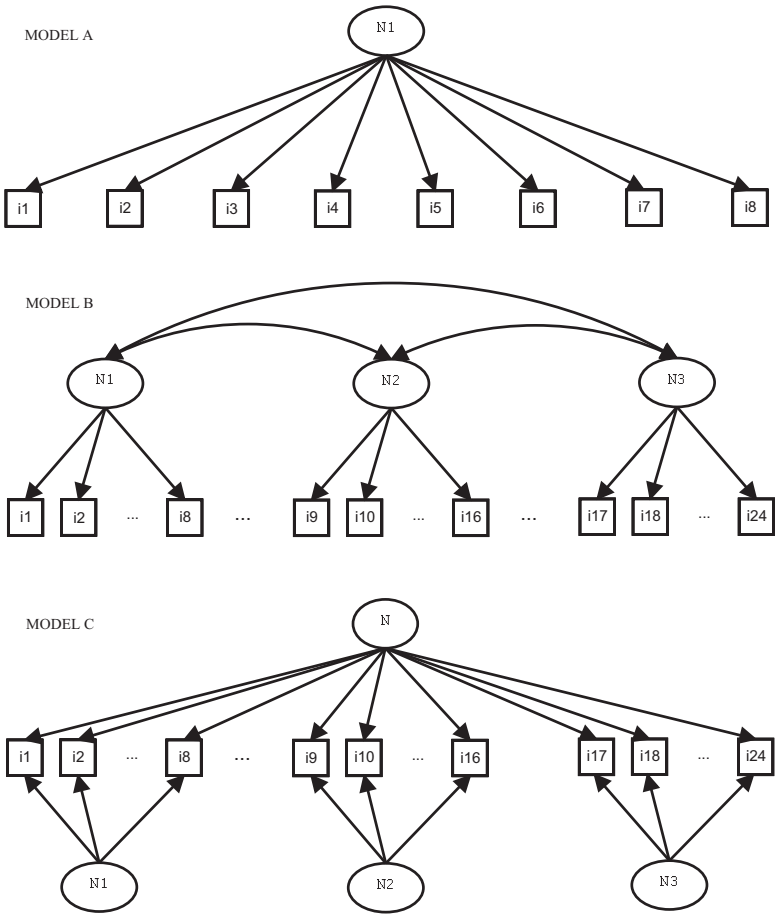

Figure 1. Representation of three different models for the Neuroticism (N) domain and three of its facets (N1 = Anxiety, N2 = Hostility, and N3 = Depression). Model A = Unidimensional; Model B: Multidimensional Correlated Traits; Model C: Bifactor. i1, . . . , i24 represent the items.

four items through UCAT provided accurate trait estimates in comparison with the complete eight-item facet scales ($r > .90$). Similar results were obtained by Nieto et al. (2017). They applied UCAT to assess each facet with four items of a new item pool based on the Big Five model and found an average correlation of $\bar{r} = .95$ between UCAT and pool facet scores. However, the application of separate UIRT models and therefore UCAT does not allow considering the intercorrelations between facets of the same domain. Consequently, the fact of ignoring such information makes UIRT inefficient to represent the Big Five personality traits.

On the other hand, MIRT based on the correlated-factors model and, by extension, MCAT based on such a model, allows studying the correlations between several factors to obtain efficient test scores (see Figure 1, Model B). Makransky et al. (2013) demonstrated that the application of MIRT improved the precision and efficiency of the NEO-PI-R facets when they were highly correlated. Thus, the facets of Neuroticism, Openness, and Conscientiousness, which showed the highest intercorrelations on average ($\bar{r} = .70$ for the former, and $\bar{r} = .60$ for the two last), obtained greater gains in precision. In addition, applying MCAT to model the facets of each domain led to facet scores as least as accurate as UIRT on average, with reductions in test length of 75% for Neuroticism, 63% for Openness, and 50% for Conscientiousness.

Although both UIRT and MIRT approaches have been applied to study the Big Five facets, they do not allow modeling simultaneously multiple hierarchically organized constructs that represent a broad trait (i.e., the domain) and several narrower subdomains (i.e., the facets). The application of the bifactor model has increased dramatically as an alternative to account for this type of construct-relevant multidimensionality of psychological measures in several fields (Reise, 2012).

MCAT-B

In the bifactor model, each item loads simultaneously on a general factor (i.e., domain) and on one of the several specific factors (i.e., facet) that account for additional common variance between clusters of items that is not explained by the general factor. All the dimensions (i.e., general and specific) are first-order factors that are assumed to be orthogonal. In Figure 1, Model C is depicted an example of a bifactor model, with a general factor representing the Neuroticism domain and three specific facets: Anxiety, Hostility, and Depression.

In personality research, several studies have applied the bifactor model to assess the Big Five traits. Chen, Hayes, Carver, Laurenceau, and Zhang (2012) illustrated the use of the bifactor model to test the multifaceted structure of the Extraversion domain of the NEO-PI-R (Costa & McCrae, 1992). Abad et al. (2016) used the bifactor model to separate the sources of variance due to the general and specific factors in each of the Big Five traits of the NEO PI-R. In addition, the application of MCAT-B has increased importantly in the last decade, mostly in the field of psychopathology, to measure multifaceted constructs such as depression, anxiety, and schizotypal personality (Gibbons et al., 2008, 2012, 2014; Gibbons, Weiss, Frank, & Kupfer, 2016; Moore, Calkins, Reise, Gur, & Gur, 2018; Sunderland, Batterham, Carragher, Calear, & Slade, 2017; Weiss & Gibbons, 2007; Zheng, Chang, & Chang, 2013). These studies have shown great savings in the number of administered items when using MCAT-B: Reductions of up to 97% were found when estimating domain scores whereas reductions ranging from 67% to 85% were found when also assessing the specific facets. In these studies, MCAT-B improved measurement precision, with CAT trait estimates being highly correlated with those obtained with the full item pool (i.e., correlations above .90). In addition, important reductions in the time required to complete the evaluations have been reported. For example, Gibbons et al. (2012) found that with a mean of 12 items, an average of 2.29 min was enough to estimate the trait level in the depression severity domain in comparison with the 51.66 min required to complete the full 389-item test.

Proposal for the Current Study

Taking all of the above into account, we propose that applying MCAT-B might provide a more suitable approach to assess the Big Five because its key feature includes modeling simultaneously the variance due to each broad domain and its narrower facets. To our knowledge, the performance of MCAT-B has not been compared with UCAT and MCAT based on correlated traits to assess the Big Five model. In addition, proposed short fixed-length versions of large Big Five inventories neither have been compared to different MCAT procedures. Thus, the main aim of this study is to assess whether a MCAT-B can provide more efficient estimates of the Big Five personality traits than three other competing approaches: a short scale, UCAT, and MCAT with correlated factors. In addition, we study whether benefits of applying MCAT-B depends on the degree of multidimensionality of the measured Big Five trait: It is expected that the bifactor model will be more advantageous with highly multidimensional traits, whereas the unidimensional approach will be preferred for traits with a strong general factor. Therefore, a secondary goal is to examine whether item responses to the Big Five personality traits are sufficiently unidimensional to apply UIRT methods instead of bifactor and other MIRT models.

Method

Participants and Procedure

The dataset includes responses from 826 undergraduate psychology students (696 women [70.91%], 175 men [20.09%]) to a pool with 360 personality items to evaluate the Big Five traits. Participants' ages ranged from 17 to 50 years ($M = 20.06$, $SD = 3.73$). For some analyses, the whole sample was randomly divided into two data sets with equal size ($n_1 = n_2 = 413$), one for model-derivation analysis and the other one for cross-validating statistical results. Participants completed the items in a psychology faculty within an official system of data collection whose purpose was the participation of students in research projects in exchange for academic compensation. The University Research Ethics Committee granted approval for the present study.

Instruments

Personality item pool. The pool is composed of 360 items rated on a 5-point Likert-type scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*), measuring the Big Five and their facets: Neuroticism (anxiety, angry/hostility, depression, self-consciousness, impulsiveness, and vulnerability), Extraversion

(warmth, gregariousness, assertiveness, activity, excitement-seeking, and positive emotions), Openness (fantasy, aesthetics, feelings, actions, ideas, and values), Agreeableness (trust, straightforwardness, altruism, compliance, modesty, and tender-mindedness), and Conscientiousness (competence, order, dutifulness, achievement striving, self-discipline, and deliberation). Statements are written in the Spanish language.

Details of the original validation of the pool are provided in Nieto et al. (2017). The items of each facet were calibrated according to the unidimensional model. Average alpha coefficients for the facets within each domain ranged from .85 (Agreeableness) to .90 (Extraversion). Within the UIRT framework, the standard error (*SE*) for trait levels θ between −3 and 3 was lower than .50 for all the facets except for Compliance (Agreeableness) and Dutifulness (Conscientiousness), which is approximately equivalent to a reliability coefficient of .75. The analysis of the internal structure using pool facet scores revealed that the items were properly designed to measure the Big Five factors of personality. The pool also showed excellent convergent validity with the NEO-FFI-3 scales, with correlations ranging from .80 to .88.

NEO-FFI-3. An external measure, the NEO-FFI-3, was included in order to examine the convergent validity of the item pool calibrated according to the bifactor model. The NEO-FFI-3 is a 60-item version of the NEO-PI-3, which is in turn a revision of the NEO PI-R, that provides measures for the Big Five domains of personality. Because Spanish versions of the NEO-PI-3 and the NEO-FFI-3 questionnaires are not available, 59 of the 60 items of the NEO-FFI-3 were selected from the Spanish version of the NEO-PI-R (Cordero, Pamos, & Seisdedos, 2008). The remaining item was translated into Spanish from the English version of the NEO-FFI-3.

Data Analysis

Calibrating each domain separately: Application of IRT bifactor model. First, the missing data rate was analyzed at the item level in the whole data set. Then, the model-derivation sample ($n_1 = 413$) was used to estimate separate exploratory bifactor graded response models (Gibbons et al., 2007) for each personality domain: A structure with a general factor representing the domain and as many specific factors as facets was specified. The Metropolis-Hastings Robbins-Monro algorithm (MHRM; Cai, 2010a, 2010b) was used for parameter estimation. The MHRM method allows missing item responses. To identify each model, marker items (i.e., those with the highest factor loading on their corresponding facet according to the unidimensional model) were specified to load only on their corresponding specific factor and on the general factor, whereas the remaining items were allowed to load on all the factors. With regard to the nonmarker items, minimally informative normal prior distributions N (0, .10) were specified for the slopes of the facets on which they theoretically should not load. Then, items with factor loadings below .20 on the general factor were excluded in an iterative procedure. At the end of this process, facets with less than five items were excluded from the analysis.

Subsequently, the cross-validation sample ($n_2 = 413$) was used to test the model previously estimated for each domain. Five fit indices were obtained for model evaluation: the M_2^* statistic for polytomous data (Cai & Hansen, 2013), the root mean square error of approximation (RMSEA) as calculated from the M_2^* values (Maydeu-Olivares, Cai, & Hernández, 2011), the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the standardized root-mean-square residual (SRMSR). CFI and TLI values of .90 or greater indicate acceptable fit, and values of .95 or greater represent good fit. For the RMSEA and SRMSR indices, values between .05 and .08 are indicative of acceptable model fit, whereas values below .05 represent good fit (Hu & Bentler, 1999; McDonald & Ho, 2002). Finally, the item parameters of each model were estimated in the whole sample ($N = 826$).

Degree of essential unidimensionality of the domains. Two bifactor-specific indices were computed: the explained common variance (ECV) and the proportion of uncontaminated correlations (PUC). The ECV (Sijtsma, 2009; Ten Berge & Sočan, 2004) reflects the common variance due to the general factor and can be easily calculated. For example, for a bifactor model with six specific factors (i.e., one per facet):

$$ECV = \frac{\sum \lambda_G^2}{\sum \lambda_G^2 + \sum \lambda_{s_1}^2 + \sum \lambda_{s_2}^2 + \sum \lambda_{s_3}^2 + \sum \lambda_{s_4}^2 + \sum \lambda_{s_5}^2 + \sum \lambda_{s_6}^2} \quad (1)$$

where λ_G are the factor loadings of the general factor and λ_{s_1} to λ_{s_6} are the factor loadings of the specific factors s_1 to s_6. High ECV values (e.g., above .85 or .90), indicate a strong general factor, so that data can be considered essentially unidimensional and therefore modeled using UIRT without seriously biasing parameter estimates. Values below .70 reflect that data are sufficiently multidimensional and MIRT models should be applied (Quinn, 2014; Stucky & Edelen, 2014).

Reise, Scheines, Widaman, and Haviland (2013) and Bonifay, Reise, Scheines, and Meijer (2015) showed that the ECV is related to parameter bias and that the PUC is an important moderator in this relationship. The PUC (Bonifay et al., 2015; Reise et al., 2013) indicates the proportion of between-item correlations that, according to the theoretical model, are not affected by the specific factors. For each Big Five domain, the PUC was computed according to its theoretical independent cluster structure. For the previous example with six facets, the PUC can be calculated as:

$$\frac{J_G \times (J_G - 1) - \sum_{p=1}^{6} J_{s_p} \times (J_{s_p} - 1)}{J_G \times (J_G - 1)} \quad (2)$$

where J_G is the total number of items of the domain and J_{s_1} to J_{s_6} are the number of items of the specific factors s_1 to s_6. Following the authors previously mentioned, as the PUC increases, the ECV becomes less important to determine the extent of parameter bias. In general terms, when the PUC is very high (e.g., $>$ .90), even low ECV values can yield unbiased parameter estimates (e.g., Reise, 2012). Rodríguez, Reise, and Haviland (2016a) suggested that when both ECV and PUC are $>$.70, low parameter bias is found.

To quantify the parameter distortion resulting from fitting multidimensional (bifactor) data to a unidimensional model, the relative bias (RB) was computed for each item as the difference between the loading on the one-factor model and the general loading on the bifactor model divided by the general factor loading on the bifactor model (Rodríguez, Reise, & Haviland, 2016b).

Then, for each domain, the overall RB was computed as the average of the individual RBs in absolute value for the items of the domain. Values below 10–15% indicate minor bias (Muthén, Kaplan, & Hollis, 1987).

Precision and evidence for convergent validity of pool scores. The alpha coefficient (α) was obtained to assess the precision of the domain and facet pool raw scores. Then, within the framework of bifactor MIRT, the multidimensional maximum a posteriori (MAP; Bock & Aitkin, 1981) method was used to obtain the trait estimates (θ) for examinees in the domains and their facets. The precision of θ estimates was evaluated with the associated standard errors.

In addition, evidence for convergent validity was obtained by computing the Pearson correlation coefficients (r) between the pool raw scores on the domains and the NEO-FFI-3 raw scores.

Post hoc simulation study. A post hoc simulation study (i.e., drawing simulees' responses from the real data) was carried out to compare, for each of the Big-Five traits, the performance of four procedures: a fixed-length short scale, UCAT, MCAT, and MCAT-B. Because several items were excluded from the initial 360-item pool in the previous calibration step, only the responses to the final 307-item pool were used to simulate the four methods. As omissions are not allowed in CAT, each examinee's response pattern was completed using item and person parameter estimates obtained in the previous calibration step with the bifactor model.

The MCAT-B was based on the bifactor model. The items were adaptively selected according to the D-Optimality criteria (i.e., maximize the determinant of the information matrix for trait estimates; see Seo & Weiss, 2015). For selecting the first item, traits (θ) were initialized to zero and from there on, MAP θ estimates ($\hat{\boldsymbol{\theta}}$) were computed according to the respondent answers. CAT stopped when 12 items were administered. For each MCAT-B, $\hat{\boldsymbol{\theta}}$ estimates were obtained for one general and several specific factors. It must be noted that, in the bifactor model, the specific factors reflect the residual variance after subtracting the effects of the general domain. That is, they reflect whether the examinee facet score is above or below the expected score after controlling for the general factor (DeMars, 2013). Previous research has suggested that removing domain-level variance may dramatically alter the meaning of the facet-level constructs, so that this residualized facet scores may have an ambiguous meaning (e.g., Simms, Prisciandaro, Krueger, & Goldberg, 2012). For this reason, and for ease of interpretation, the expected or predicted observed scores on the facets (τ_f) and the domains (τ_d), which reflect the respondent's overall standing on each scale, were obtained. For example, the expected score of a respondent in facet f was obtained as the sum of the expected scores on the items measuring it (DeMars, 2013):

$$\tau_f = \sum_{j \in f} \sum_{k=1}^{K} k P_{jk}(\hat{\boldsymbol{\theta}}) \tag{3}$$

where k runs from 1 to K, the number of response categories, and $P_{jk}(\hat{\boldsymbol{\theta}})$ is the probability for a respondent with a $\hat{\boldsymbol{\theta}}$ estimate of selecting response category k of item j.

For the UCAT and the MCAT, the same CAT specifications were used, but based on the UIRT and the MIRT models, respectively. Thus, to apply these procedures, data were calibrated separately for each domain according to the UIRT (i.e., one general factor for all the items in the domain) and the MIRT models (i.e., one factor per facet). Again, for comparability with the MCAT-B, expected scores were obtained based on $\hat{\boldsymbol{\theta}}$ estimates. Finally, for each domain, a fixed-length short scale was developed with the 12 items with the highest factor loadings on the UIRT model. Expected scores were again obtained, based on $\hat{\boldsymbol{\theta}}$ estimates.

The performance of the simulated tests was examined according to two aspects: (a) accuracy and (b) item pool usage. Test accuracy was examined with the correlation between the pool raw scores and the expected scores on the tests. Pool raw scores on a domain/facet were obtained by summing the raw responses in the items of the domain/facet. Item pool usage of the tests was calculated for each facet as the proportion of items belonging to the facet that was administered to the total of simulees.

Evidence for convergent and discriminant validity of the methods. First, for each simulated 12-item test, evidence for convergent validity was obtained by computing the correlations between the expected scores on the domains and the NEO-FFI-3 raw scores. Second, as the multidimensional procedures allow to estimate the facet scores, the intercorrelations between the expected scores on the facets were obtained for the MCAT and MCAT-B methods. For each procedure and domain, the within-domain convergent correlations between facet (expected) scores on the same domain, and the between-domain discriminant correlations between the facet (expected) scores on the domain and the facet (expected) scores on the remaining domains, were analyzed to obtain evidence for convergent and discriminant validity, respectively. Because the convergent correlations between facets of the same domain are expected to be positive in all the cases, the average value was reported. Besides, as the discriminant correlations may take positive or negative values depending on the facets involved, the average absolute value was computed in this case. The convergent and discriminant correlations between the pool raw scores on the facets were also obtained so as to establish a baseline for comparisons.

All the statistical analyses were performed using the R (R Core Team, 2017) package *mirt* (Chalmers, 2012). The program with the CAT algorithms was developed with the package *mirtCAT* (Chalmers, 2016).

Results

Calibrating Each Domain Separately: Application of IRT Bifactor Model

Missing data rate for item nonresponse was very low, with a maximum value of 2%. A total of 53 out of the 360 items in the pool were excluded because they presented factor loadings below .20 on the general factor of their correspondent model. The largest number of excluded items was 18 both for Neuroticism and Conscientiousness domains. It should be mentioned that, in the case of Neuroticism, the Impulsiveness facet was excluded because it had less than five items after the item selection analysis. In relation to the remaining domains, five items were excluded in the case of Extraversion and six both for Openness and Agreeableness traits. The final pool was composed of 307 items, with an average number of 61 items per personality domain.

Table 1 shows the goodness-of-fit statistics for the bifactor solutions. Model fit for the five domains was excellent; that is, in general, all the indices had values according to the recommended

Table 1
Goodness of Fit Statistics for the Five IRT Bifactor Models in the Cross-Validation Sample ($n_2 = 413$)

Domain	M_2^*	*df*	RMSEA	TLI	CFI	SRMSR
Neuroticism	1,818.52	1,151	.04	.95	.96	.05
Extraversion	2,028.48	1,406	.03	.96	.97	.06
Openness	2,448.04	1,688	.03	.94	.95	.06
Agreeableness	1,821.12	1,301	.03	.96	.97	.05
Conscientiousness	1,479.74	922	.04	.95	.96	.07

Note. M_2^* = fit statistic for polytomous data of Cai and Hansen (2013); *df* = degrees of freedom of M_2^*; RMSEA = root mean square error of approximation; TLI = Tucker-Lewis index; CFI = comparative fit index; SRMSR = standard root mean square residual.

criteria for good fit. Average values for the indices were: $\overline{\text{CFI}}$ = .96, $\overline{\text{TLI}}$ = .95, $\overline{\text{RMSEA}}$ = .03, $\overline{\text{SRMSR}}$ = .06.

In the final bifactor solutions, all the item parameter estimates for the corresponding theoretical structure were significantly different from zero ($p < .05$). Table 2 shows the average item loadings on the general and specific factors for the five domains. The average item loadings on the general factor ranged from .43 (Agreeableness and Conscientiousness) to .51 (Extraversion). Regarding the specific factors, average item loadings ranged from .25 to .48 for Neuroticism, from .30 to .48 for Extraversion, from .18 to .62 for Openness, from .22 to .55 for Agreeableness, and from .23 to .69 for Conscientiousness. For the five bifactor solutions, the average cross-loading in absolute value was low (.04 in all the cases).

Degree of Essential Unidimensionality of the Domains

The ECV, PUC, and RB values for the five bifactor solutions are presented in Table 3. The average ECV for the five domains was .52. This indicates that, overall, the general factor explains about 52% of the common variance, whereas approximately 48% of the common variance is distributed across the specific factors in the five domains. Extraversion showed the highest value (ECV = .62), whereas Conscientiousness yielded the lowest (ECV = .44). The average PUC was .83, which indicates that the great majority of the correlations theoretically reflect the general factor in the five domains. Regarding the RB, only Extraversion showed low parameter bias (RB = 7%). For Neuroticism, the RB was 10%, indicating non-negligible bias. Parameter bias was severe in the case of Openness (RB = 16%), Conscientiousness (RB = 16%), and Agreeableness (RB = 19%). It should be noted that lower RB values were associated with higher ECV values. For example, for Extraversion, which showed the highest ECV value, the RB was minor.

Precision and Evidence for Convergent Validity for Pool Scores

The alpha coefficient for the pool scores on the domains was excellent, with values that ranged from .92 (Conscientiousness) to .95 (both for Neuroticism and Extraversion). Both for Openness and Agreeableness, α was .93. Regarding the precision of the pool facet scores, almost all alpha values were above .70, except for the case of compliance (α = .67) and dutifulness (α = .60) facets.

Table 2
Bifactor Models for the Big Five Domains: Number of Final Items and Average Item Loadings on the General and Specific Factors

Domain/facets	Number of final items	Average item loadings: General factor	Average item loadings: Specific factor
Neuroticism	58	.50	
Anxiety	11	.53	.25
Angry/hostility	9	.37	.48
Depression	12	.64	.33
Self-consciousness	14	.41	.43
Vulnerability	12	.52	.43
Extraversion	64	.51	
Warmth	13	.53	.31
Gregariousness	14	.53	.30
Assertiveness	9	.54	.38
Activity	11	.47	.30
Excitement seeking	7	.41	.48
Positive emotions	10	.57	.47
Openness	69	.45	
Fantasy	13	.39	.62
Aesthetics	13	.51	.40
Feelings	9	.38	.48
Actions	13	.43	.48
Ideas	11	.56	.18
Values	10	.39	.45
Agreeableness	62	.43	
Trust	12	.29	.55
Straightforwardness	9	.39	.49
Altruism	12	.61	.25
Compliance	8	.40	.33
Modesty	10	.31	.52
Tender-mindedness	11	.52	.22
Conscientiousness	54	.43	
Competence	8	.40	.53
Order	11	.42	.55
Dutifulness	5	.40	.38
Achievement striving	11	.45	.33
Self-discipline	11	.54	.23
Deliberation	8	.32	.69

Values for the facets of each domain ranged from .81 (angry/hostility) to .91 (depression) for Neuroticism, from .79 (excitement seeking) to .90 (positive emotions) for Extraversion, from .76 (feelings) to .91 (fantasy) for Openness, from .67 (compliance) to .86 (altruism) for Agreeableness, and from .60 (dutifulness) to .89 (order, deliberation) for Conscientiousness.

Table 3
Explained Common Variance (ECV), Percentage of Uncontaminated Correlations (PUC), and Relative Bias (RB) for the Bifactor Models

Domain	ECV	PUC	RB (%)
Neuroticism	.58	.81	10
Extraversion	.62	.84	7
Openness	.46	.84	16
Agreeableness	.50	.84	19
Conscientiousness	.44	.84	16

Figure 2 illustrates the *SE* for the IRT θ estimates in the general domains of the bifactor solutions when the complete pool is administered. For trait estimates between −3 and 3, the *SE* was lower than .40 for the five domains, which is approximately equivalent to a reliability coefficient of .84. On average, the lowest *SE*s were for Extraversion ($\overline{SE}$ = .26) and Neuroticism ($\overline{SE}$ = .27) whereas Conscientiousness showed the largest value ($\overline{SE}$ = .34). For Openness and Agreeableness, the $\overline{SE}$ was .32. This indicates that the item pool calibrated according to the bifactor model provides excellent information across the different trait levels of each domain.

Regarding the convergent validity between the pool scores on the domains and scores on the NEO-FFI-3 scales, the degree of association was excellent for the five traits. Neuroticism and Extraversion showed the highest values (in both cases, r = .90) whereas the lowest values were for Agreeableness and Conscientiousness (r = .83). For Openness, r = .85.

Post-Hoc Simulation Study

Table 4 shows the correlations between the pool raw scores and IRT expected scores on the domain and facets for the four simulated tests. At the domain level, the multidimensional tests (e.g., MCAT and MCAT-B) showed the best performance with the highest correlations on average ($\bar{r}$ = .94 and .93, respectively), whereas the unidimensional procedures (e.g., short scale and UCAT) were generally less accurate ($\bar{r}$ = .89 for both methods). The MCAT and MCAT-B tests performed similarly across the five domains (e.g., for Neuroticism, r = .94 for both methods), and UCAT and the short scale showed similar results (e.g., for Extraversion, r = .95 for the two tests). Taking this into account, the results for each domain are summarized by comparing the correlations of the MCAT-B and UCAT procedures. Both tests showed statistically significant differences ($p < .001$) in performance in favor of MCAT-B for Agreeableness, Openness, Conscientiousness, and Neuroticism ($r_{MCAT-B} - r_{UCAT}$ = .09, .05, .05, and .04, respectively). Only in the case of Extraversion did both tests perform similarly. These differences in performance are consistent with previous results regarding the essential unidimensionality of the domains. Thus, Conscientiousness, Openness, and Agreeableness, which showed the lowest ECVs (.44, .46, and .50, respectively), also presented the highest parameter biases when a unidimensional model was fit to the data (RB = 16%, 16%, and 19%, respectively) and, therefore, the highest differences in performance between MCAT-B and UCAT. In the case of Extraversion, this domain presented the highest ECV (.62) and the lack of differences between UCAT and MCAT-B is, in turn, consistent with the slight bias (RB = 7%) found when a one-factor model was fit to the data.

On another hand, at the facet-level, the MCAT and the MCAT-B procedures revealed a similar performance on average across the five domains: for Neuroticism $\bar{r}_{MCAT}$ = .88 and $\bar{r}_{MCAT-B}$ = .87, for Extraversion $\bar{r}_{MCAT}$ = .89 and $\bar{r}_{MCAT-B}$ = .88,

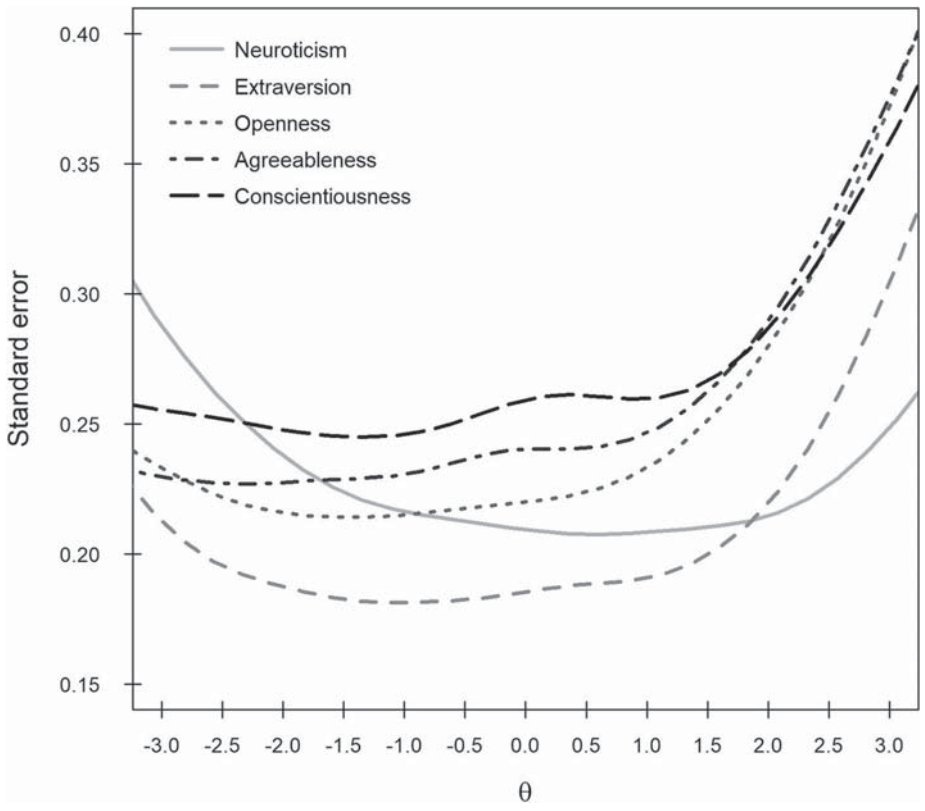

Figure 2. Standard error across pool domain scores for the Big Five traits.

Table 4
Pearson Correlations Between the Pool Domain/Facet Scores and Expected Scores on the Big Five Domain and Facets for the Short Scale, UCAT, MCAT, and MCAT-B

Domain/facet	Short scale	UCAT	MCAT	MCAT-B
Neuroticism	.89	.90	.94	.94
Anxiety			.86	.89
Angry/hostility			.86	.84
Depression			.93	.91
Self-consciousness			.86	.85
Vulnerability			.90	.88
Extraversion	.95	.95	.96	.95
Warmth			.90	.89
Gregariousness			.86	.87
Assertiveness			.91	.89
Activity			.83	.83
Excitement seeking			.90	.90
Positive emotions			.92	.92
Openness	.89	.88	.94	.93
Fantasy			.89	.88
Aesthetics			.90	.89
Feelings			.80	.80
Actions			.85	.87
Ideas			.85	.83
Values			.84	.84
Agreeableness	.85	.83	.93	.92
Trust			.86	.85
Straightforwardness			.85	.84
Altruism			.89	.88
Compliance			.79	.78
Modesty			.84	.83
Tender-mindedness			.84	.85
Conscientiousness	.87	.88	.93	.93
Competence			.87	.85
Order			.90	.92
Dutifulness			.78	.78
Achievement striving			.85	.85
Self-discipline			.85	.84
Deliberation			.91	.90

Note. UCAT = Unidimensional Computerized Adaptive Test; MCAT = multidimensional computerized adaptive testing; MCAT-B = multidimensional computerized adaptive testing based on the bifactor model.

for Openness $\bar{r}_{MCAT} = .86$ and $\bar{r}_{MCAT-B} = .85$, for Agreeableness $\bar{r}_{MCAT} = \bar{r}_{MCAT-B} = .84$, and for Conscientiousness $\bar{r}_{MCAT} = \bar{r}_{MCAT-B} = .86$.

Figure 3 shows the percentage of items belonging to each facet that was administered in each simulated test to assess the domains. In the case of the short scale, all the respondents answered the same items, which were those with the highest loadings when applying UIRT. As the items are selected according to their loading on the one-factor model, there are a different number of items for each facet, and sometimes, a facet is not even measured in the short scale. The same thing occurred with UCAT because the most informative items are selected. This explains the heterogeneous representation of the facets across the five domains for the short scales and UCAT. Indeed, the facets with the highest percentages of representation were the same when using the short scale and UCAT. On the contrary, in the case of MCAT and MCAT-B, all the facets were represented to a similar degree. For example, in the case of Extraversion, the percentage of items belonging to each facet ranged from 13% to 20% in the MCAT and from 13% to 24% in the case of the MCAT-B. This indicates that the multidimensional approaches provide a better content-balance strategy than the unidimensional ones. It should be noted that, in the case of Extraversion, which was shown to be the most unidimensional domain, the distributions for the short scale and the UCAT tended to be more uniform; that is, more similar to the distributions of the multidimensional tests than were observed in the remaining domains.

Evidence for Convergent and Discriminant Validity of the Methods. The results for the convergent validity with the NEO-FFI-3 scales are shown in Table 5. For Agreeableness, which proved to be one of the most multidimensional constructs (i.e., the one which the highest RB), the multidimensional procedures showed stronger convergence (e.g., $r_{\text{MCAT}} = .79 > r_{\text{UCAT}} = .65$). In contrast, in the case of Extraversion, which was the most unidimensional domain, the four procedures showed slight differences in performance (i.e., the greater difference was .02). For the remaining domains, the differences between tests were also small (i.e., the greater difference was .05) and the evidence was mixed. For Neuroticism and Openness, the unidimensional procedures showed stronger convergence than the multidimensional procedures (e.g., for Neuroticism, $r_{\text{UCAT}} = .90 > r_{\text{MCAT-B}} = .86$), whereas for Conscientiousness all the CATs showed similar performance and better convergence than the short test (e.g., $r_{\text{short}} = .74 < r_{\text{MCAT-B}} = .79$).

The results of the analysis of the convergent and discriminant correlations between facets for the MCAT, the MCAT-B, and the item pool are shown in Table 6. Regarding the within-domain convergent correlations (Table 6, top), they were systematically higher on average for the facets of those domains that proved to be more unidimensional and lower for the facets of the domains that showed a more multidimensional structure (e.g., with the pool raw scores, the highest average correlation was .53 for Extraversion, whereas the lowest was .32 for Conscientiousness). Regarding the methods, both multidimensional tests produced an overestimation of the correlations that was slightly higher in the case of the MCAT-B.

As expected, the discriminant correlations (Table 6, bottom) were lower than the convergent correlations (e.g., for Extraversion, the average absolute discriminant r for the MCAT-B was .21 whereas the average convergent r was .72). This indicates that the facets of a domain were well differentiated from the facets of other domains. Both the MCAT and the MCAT-B performed similarly across the five domains.

Discussion

The purpose of this study was to examine whether a MCAT-B can more efficiently provide estimates of the Big Five traits than three other competing approaches: UCAT, MCAT with correlated factors, and a short scale. For the five domains, the estimated bifactor model with a general factor representing the domain and several specific factors representing the corresponding facets fit the data well. In addition, convergent validity between the calibrated pool and the NEO-FFI-3 questionnaire was excellent for the five domains. When the essential unidimensionality of the domains was tested, the PUC was high in all the cases, that is, the influence of the specific facets was low in the factor structure, but the ECV did not suggest the presence of a sufficiently strong general factor. Extraversion obtained the ECV value closest to .70 (ECV = .62),

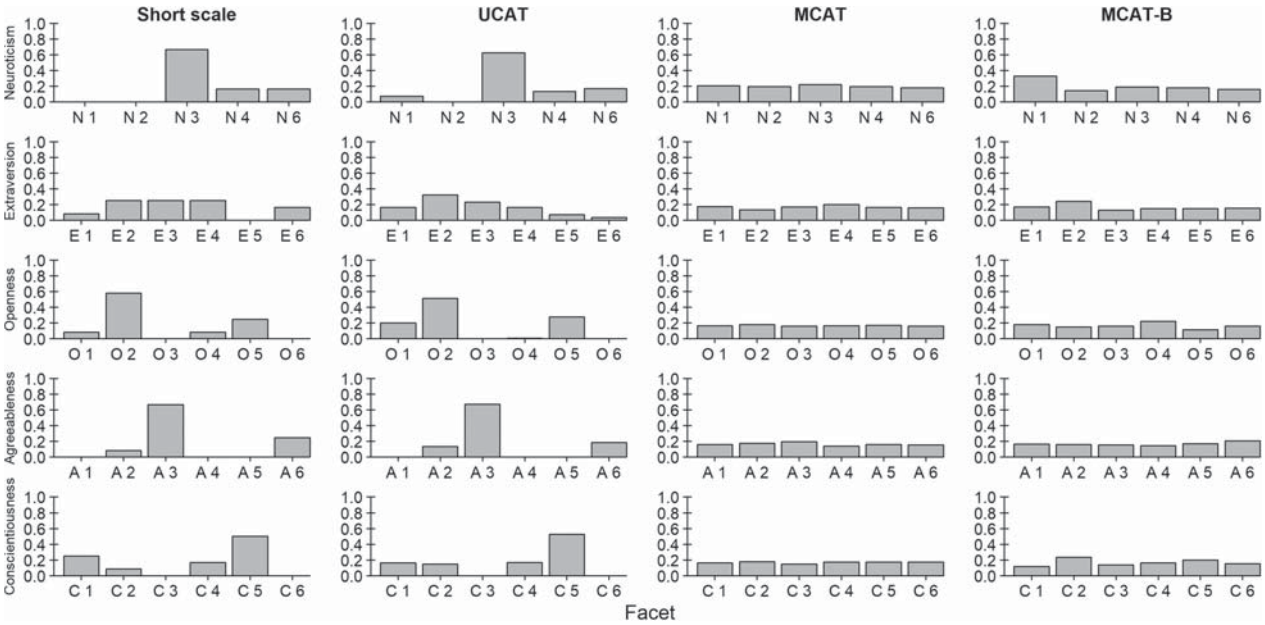

Figure 3. Rate of items selected from each specific facet in the four types of tests for each personality domain. UCAT = Unidimensional Computerized Adaptive Test; MCAT = Multidimensional Computerized Adaptive Test; MCAT-B = Multidimensional Computerized Adaptive Test with bifactor model; N1, . . . , N6: Facets of Neuroticism; E1, . . . , E6: Facets of Extraversion; O1, . . . , O6: Facets of Openness; A1, . . . , A6: Facets of Agreeableness; C1, . . . , C6: Facets of Conscientiousness.

closely followed by Neuroticism (ECV = .58). Although both domains showed similar ECV values, the parameter bias was higher for Neuroticism. The remaining domains showed severe parameter bias, with RB values greater than or equal to 16% when UIRT was applied. Taking this into account, none of the domains clearly showed a strong unidimensional structure. However, Extraversion was the one that obtained the lowest parameter bias, so it is the only domain whose structure is closer to unidimensionality.

The results from the post hoc simulation study revealed that, generally, for each domain, the unidimensional instruments (i.e., the short scale and UCAT) showed a similar performance, worse than did the multidimensional procedures (i.e., MCAT and MCAT-B). Specifically, results for each domain were closely related to its degree of essential unidimensionality. Thus, only in the case of Extraversion, which was the most unidimensional domain, the short scale and UCAT were shown to be as efficient as the multidimensional procedures in recovering the domain scores. Regarding the pool usage of UCAT for the five domains, there was a greater representation of the facets with a greater number of highly informative items, whereas few or no items were selected for the remaining facets. This is consistent with the representation of the facets for the short scales, which were composed of the 12 best items in the UIRT model (i.e., the items with the highest factor loadings). These results are in line with the findings of Reise and Henson (2000), who concluded that similar results can be found using UCAT and the best items (i.e., the most informative) of a scale, although they referred to the unidimensional evaluation of the Big Five facets. It should be noted that, only for Extraversion, the distributions for the short scale and the UCAT tended to be more uniform; that is, both instruments tended to better balance the representation of facets. Despite this, for the

Table 5
Convergent Validity With the NEO-FFI-3 for the Short Scale, UCAT, MCAT, and MCAT-B

Test	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
Short scale	.89	.87	.83	.65	.74
UCAT	.90	.88	.85	.65	.78
MCAT	.87	.87	.82	.79	.77
MCAT-B	.86	.86	.81	.77	.79
Item pool	**.90**	**.90**	**.85**	**.83**	**.83**

Note. The values for the 307-item pool are shown in boldface. UCAT = Unidimensional Computerized Adaptive Test; MCAT = multidimensional computerized adaptive testinig; MCAT-B = multidimensional computerized adaptive testing based on the bifactor model.

Table 6
Convergent and Discriminant Correlations for the Item Pool, MCAT, and MCAT-B

Test	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
Average within-domain convergent correlation					
MCAT	.63	.69	.49	.59	.46
MCAT-B	.67	.72	.54	.59	.52
Item pool	**.51**	**.53**	**.35**	**.41**	**.32**
Average absolute between-domain discriminant correlation					
MCAT	.20	.22	.12	.12	.15
MCAT-B	.20	.21	.10	.12	.14
Item pool	**.20**	**.22**	**.13**	**.14**	**.17**

Note. The average within-domain convergent correlation refers to the average value of the individual correlations between the facet (expected) scores on the same domain. The average absolute between-domain discriminant correlation refers to the average value of the individual correlations (in absolute value) between the facet (expected) scores on one domain and the facet (expected) scores on the remaining domains. The values for the 307-item pool are shown in boldface. MCAT = multidimensional computerized adaptive testing; MCAT-B: multidimensional computerized adaptive testing based on the bifactor model.

UCAT, the content balance remained disproportionate in favor of some facets (e.g., gregariousness) and the short scale did not contain any item from the facet of excitement seeking. Misrepresentation of facets has been targeted as a limitation of the use of short scales because it can constitute a source of model misfit (Gignac et al., 2007). In the case of UCAT, the item pool usage could be improved by setting content constraints for the specific facets (Makransky et al., 2013). In this regard, the MCAT and MCAT-B methods showed a clear advantage in terms of balancing pool usage not only for Extraversion but for all the Big Five domains, so that items from all the facets were always administered in similar proportions.

For the domains that proved to be more multidimensional (Neuroticism, Agreeableness, Openness, and Conscientiousness), the MCAT and MCAT-B methods outperformed the unidimensional procedures when estimating the domain scores. Regarding the recovery of the pool facet scores, both procedures showed a similar performance across the five domains. Besides, according to the evidence of validity obtained in this study, at the domain level, the multidimensional methods presented greater validity for the Agreeableness domain, which showed to be the most multidimensional according to the RB. For the remaining domains, the differences between procedures were generally small and, in some cases (i.e., for Neuroticism and Openness) favored the unidimensional methods. This unexpected advantage might be because the criteria (i.e., the NEO-FFI scores) are brief measures that are directly designed to measure the domain factors, as is the case of the short test and the UCAT. In contrast, the goal in a multidimensional CAT is to recover not only the domain scores but also the facet scores. In this sense, whereas the better content balance of multidimensional tests led to a better recovering of the pool raw score, it might also be slightly reducing the efficiency for measuring the general domain. Consistent with this, the differences between procedures were smaller for the Extraversion domain, in which the unidimensional tests achieved a good content balance.

At the facet level, both multidimensional methods performed similarly and only showed slight differences in the convergent and discriminant correlations. It must be noted that some inflation was found for the within-domain convergent correlations (i.e., between facets of the same domain). This overestimation might be partly due to the bias in the Bayesian estimates produced by the inclusion of the prior correlation matrix, and thus caution should be exercised when interpreting these correlations between estimates (Segall, 1996).

Despite the similarities between the multidimensional tests, the use of bifactor modeling offers several advantages over the correlated-factor model, which make it a more desirable approach to assess multifaceted constructs. It should be noted that this advantages are not inherent to CAT. First, as we have illustrated in this study, some bifactor-derived indices (i.e., ECV and PUC) can be easily obtained by researchers to examine the degree of unidimensionality of the constructs to determine whether a UIRT or MIRT model is required (e.g., Rodríguez et al., 2016a, 2016b). Second, the bifactor model yields an estimated score in the general domain with an associated standard error, which is an indicator of the accuracy of the overall measure. Although in MIRT with correlated factors, a general score in the domain can be obtained by averaging the results over the specific facets (e.g., Makransky et al., 2013), this cannot be directly estimated and, therefore, the model does not provide any information on its accuracy. A third advantage not explored here is that it also allows estimating the accuracy associated with the IRT facet scores. Although this residualized facet scores are difficult to interpret, many studies have shown how they can contribute to the incremental prediction of several psychological measures above and beyond scores on the general factor. For example, McAbee et al. (2014) applied separate bifactor models for each of the six traits of the HEXACO model of personality (Lee & Ashton, 2004) and examined the role of the general factor and the specific facet scores for predicting students' performance. They concluded that modeling facet scores enables researchers to explain interesting but complex relations between narrow personality traits and student performance outcomes, which cannot be otherwise studied. In addition, facet scores may be especially informative for assessment contexts where individual personality profiles need to be developed attending not only to the broader trait (i.e., the domain) but also to the individual differences reflected by the narrower traits (i.e., the facets). Related to this, it is important to note that several authors have highlighted the importance of evaluating whether subscale scores in multidimensional measurement models have added value over the total score and, therefore, if they should be computed and used (e.g., Reise et al., 2013; Sinharay, 2013; Sinharay, Puhan, & Haberman, 2011).

In the case of the bifactor model, some indices that can be applied to subscales, such as omega and omega hierarchical, have been used for this purpose (Rodríguez et al., 2016a, 2016b).

Besides, a number of studies have reported great gains in efficiency associated with the use of MCAT-B when estimating domain and facets scores (e.g., Gibbons et al., 2012). In this study, the time required to complete the final 307-item pool was 62.23 min approximately and, proportionally, about 12.16 min to complete any of the 60-item adaptive versions. This supposes important reductions of testing time and test length (i.e., 83%). Moreover, considering that in this short time the MCAT-B procedure provides both the domain and facet scores, the advantage over UCAT (and the short scale), which only provides the domain score, is evident when facet scores are required (e.g., for diagnosis purposes). Likewise, as mentioned above, in such time the use of the bifactor model allows to obtain a measure of precision of the domain score (i.e., the *SE*) which cannot be obtained with the MCAT procedure. This is especially relevant when evaluating multidimensional constructs and the objective of the evaluation is to provide the domain score. For example, Moore et al. (2018) applied the bifactor model to design a CAT to measure the general trait of schizotypal personality, which includes several features or dimensions (e.g., cognitive-perceptual). As these authors pointed out, in these cases the bifactor model allows to account for multidimensionality through the inclusion of the specific factors, which in fact contribute to the measurement precision of the general domain. Moreover, fitting the unidimensional IRT model to multidimensional data would not be optimal either because it may lead to biased item parameter estimates (Reise, Moore, & Maydeu-Olivares, 2011; Reise, Cook, & Moore, 2015).

Previous studies assessing the Big Five model with MCAT (Makransky et al., 2013) or applying MCAT-B (Gibbons et al., 2008, 2012, 2014, 2016; Moore et al., 2018; Sunderland et al., 2017; Weiss & Gibbons, 2007; Zheng et al., 2013) specified confirmatory structures to calibrate the item pools. In the current study, we illustrated the application of more realistic bifactor exploratory models to measure the Big Five adaptively.

The current study has several limitations that deserve further discussion. First, we are aware of the problems of the generalizability of the findings to other contexts due to the specificity of the study sample. In this regard, examining the intercorrelations between the five personality factors, we have found they are consistent with previous research. For example, Neuroticism correlated negatively with the remaining domains and showed high associations with Extraversion ($r = -.56$) and Conscientiousness ($r = -.19$), whereas domains such as Openness and Agreeableness showed lower correlations ($r = .13$; Mount, Barrick, Scullen, & Rounds, 2005; van der Linden, te Nijenhuis, & Bakker, 2010). Therefore, although in this study, the pattern of relationships between the Big Five domains is similar to that previously found, further research is required to replicate these results in other subpopulations. Second, a post hoc simulation was conducted to examine the performance of the tests, and therefore simulees' responses were drawn from the real dataset. Although real data simulations are essential to evaluate how CAT procedures will operate with real respondents (Thompson & Weiss, 2011), it is necessary to carry out additional studies with live examinees to investigate their performance in real testing settings. Third, we have defined the adaptive algorithms according to a unique item selection criterion (i.e., D-Optimality). Future research should evaluate the performance of alternative item administration criteria. For example, Seo and Weiss (2015) showed through a Monte Carlo simulation study that the D_s-Optimality criterion worked well when the focus is on measuring the general factor of a bifactor model, whereas other rules such as D- or A-optimality improved the measurement of the specific factors.

In closing, this study provides two main contributions to previous research concerning the adaptive assessment of personality. First, the Big Five domains are essentially multidimensional constructs and, therefore, they cannot be adequately evaluated through the application of a unidimensional model. Second, and related to the previous conclusion, MCAT-B constitutes a preferential framework for adaptively assessing the Big Five of personality because it allows assessing the general domains while representing the multidimensionality due to the specificity of the facets. Several other applications of the bifactor model have been illustrated to address a number of issues of interest in personality research. In this regard, it is common to observe how individual differences in the response style constitute a source of variance that can systematically distort the factor structure of personality instruments and lead to model misfit (e.g., Podsakoff, MacKenzie, & Podsakoff, 2012). Abad et al. (2016) illustrated how the inclusion of an acquiescence method factor can be a useful tool to separate variance explained by general and specific traits of personality from variance due to the acquiescent response style. It would also be interesting to include social desirability item markers to study its relationship with the Big Five domains and facets and to determine how this response style can affect the prediction of different psychological constructs (see, e.g., Ferrando, Lorenzo-Seva, & Chico, 2009). Taking all the above into account, future research in the area of adaptive assessment of personality should be oriented toward the modeling and study of response styles during the phases of calibration and administration of the item pool.

References

Abad, F. J., Sorrel, M. A., García, L. F., & Aluja, A. (2016). Modeling general, specific, and method variance in personality measures: Results for ZKA-PQ and NEO-PI-R. *Assessment*. Advance online publication. http://dx.doi.org/10.1177/1073191116667547

Ashton, M. C., Paunonen, S. V., & Lee, K. (2014). On the validity of narrow and broad personality traits: A response to Salgado, Moscoso, and Berges (2013). *Personality and Individual Differences, 56,* 24–28. http://dx.doi.org/10.1016/j.paid.2013.08.019

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459. http://dx.doi.org/10.1007/BF02293801

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling, 22,* 504–516. http://dx.doi.org/10.1080/10705511.2014.938596

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hasings Robbins-Monro algorithm. *Psychometrika, 75,* 33–57. http://dx.doi.org/10.1007/s11336-009-9136-x

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35,* 307–335. http://dx.doi.org/10.3102/1076998609353115

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and*

Statistical Psychology, 66, 245–276. http://dx.doi.org/10.1111/j.2044-8317.2012.02050.x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1–29. http://dx.doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71,* 1–39. http://dx.doi.org/10.18637/jss.v071.i05

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality, 80,* 219–251. http://dx.doi.org/10.1111/j.1467-6494.2011.00739.x

Cordero, A., Pamos, A., & Seisdedos, N. (2008). *Revised NEO Personality Inventory (NEO PI-R) manual. Spanish adaptation*. Madrid, Spain: TEA Ediciones.

Costa, P., & McCrae, R. R. (1992). *NEO PI-R manual professional*. Odessa, FL: Psychological Assessment Resources, Inc.

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing, 13,* 354–378. http://dx.doi.org/10.1080/15305058.2013.799067

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2009). A general factor-analytic procedure for assessing response bias in questionnaire measures. *Structural Equation Modeling, 16,* 364–381. http://dx.doi.org/10.1080/10705510902751374

Forbey, J. D., & Ben-Porath, Y. S. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 Computerized Adaptive Version. *Psychological Assessment, 19,* 14–24. http://dx.doi.org/10.1037/1040-3590.19.1.14

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . . Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31,* 4–19. http://dx.doi.org/10.1177/0146621606289485

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology, 12,* 83–104. http://dx.doi.org/10.1146/annurev-clinpsy-021815-093634

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., . . . Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59,* 361–368. http://dx.doi.org/10.1176/ps.2008.59.4.361

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *JAMA Psychiatry, 69,* 1104–1112. http://dx.doi.org/10.1001/archgenpsychiatry.2012.14

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2014). Development of the CAT-ANX: A computerized adaptive test for anxiety. *The American Journal of Psychiatry, 171,* 187–194. http://dx.doi.org/10.1176/appi.ajp.2013.13020178

Gignac, G. E., Bates, T. C., & Jang, K. L. (2007). Implications relevant to CFA model misfit, reliability, and the five-factor model as measured by the NEO-FFI. *Personality and Individual Differences, 43,* 1051–1062. http://dx.doi.org/10.1016/j.paid.2007.02.024

Goldberg, L. R. (1999). A broad-band width, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, the Netherlands: Tilburg University Press.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. John, R. Robins, & L. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). New York, NY: Guilford Press.

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51,* 78–89. http://dx.doi.org/10.1016/j.jrp.2014.05.003

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research, 39,* 329–358. http://dx.doi.org/10.1207/s15327906mbr3902_8

Makransky, G., Mortensen, E. L., & Glas, C. A. (2013). Improving personality facet scores with multidimensional computer adaptive testing: An illustration with the NEO PI-R. *Assessment, 20,* 3–13. http://dx.doi.org/10.1177/1073191112437756

Maples, J. L., Guan, L., Carter, N. T., & Miller, J. D. (2014). A test of the International Personality Item Pool representation of the Revised NEO Personality Inventory and development of a 120-item IPIP-based measure of the five-factor model. *Psychological Assessment, 26,* 1070–1084. http://dx.doi.org/10.1037/pas0000004

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of IRT and factor analysis models. *Structural Equation Modeling, 18,* 333–356. http://dx.doi.org/10.1080/10705511.2011.581993

McAbee, S. T., Oswald, F. L., & Connelly, B. S. (2014). Bifactor models of personality and college student performance: A broad versus narrow view. *European Journal of Personality, 28,* 604–619.

McCrae, R. R., & Costa, P. T., Jr. (2007). Brief versions of the NEO-PI-3. *Journal of Individual Differences, 28,* 116–128. http://dx.doi.org/10.1027/1614-0001.28.3.116

McCrae, R. R., Costa, P. T., Jr., & Martin, T. A. (2005). The NEO-PI-3: A more readable revised NEO Personality Inventory. *Journal of Personality Assessment, 84,* 261–270. http://dx.doi.org/10.1207/s15327752jpa8403_05

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7,* 64–82. http://dx.doi.org/10.1037/1082-989X.7.1.64

Moore, T. M., Calkins, M. E., Reise, S. P., Gur, R. C., & Gur, R. E. (2018). Development and public release of a computerized adaptive (CAT) version of the Schizotypal Personality Questionnaire. *Psychiatry Research, 263,* 250–256. http://dx.doi.org/10.1016/j.psychres.2018.02.022

Mount, M. K., Barrick, M. R., Scullen, S. M., & Rounds, J. (2005). Higher-order dimensions of the big five personality traits and the big six vocational interest types. *Personnel Psychology, 58,* 447–478. http://dx.doi.org/10.1111/j.1744-6570.2005.00468.x

Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52,* 431–462. http://dx.doi.org/10.1007/BF02294365

Nieto, M. D., Abad, F. J., Hernández-Camacho, A., Garrido, L. E., Barrada, J. R., Aguado, D., & Olea, J. (2017). Calibrating a new item pool to adaptively assess the Big Five. *Psicothema, 29,* 390–395.

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences, 43,* 971–990. http://dx.doi.org/10.1016/j.paid.2007.03.017

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63,* 539–569. http://dx.doi.org/10.1146/annurev-psych-120710-100452

Quinn, H. O. (2014). *Bifactor models, explained common variance (ECV), and the usefulness of scores from unidimensional item response theory analyses* (Unpublished Master's thesis)., The University of North Carolina at Chapel Hill, NC.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696. http://dx.doi.org/10.1080/00273171.2012.715555

Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. Revicki (Eds.), *Handbook of item response theory modeling*. New York, NY: Routledge.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7,* 347–364. http://dx.doi.org/10.1177/107319110000700404

Reise, S. P., Moore, T., & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement, 71,* 684–711. http://dx.doi.org/10.1177/0013164410378690

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73,* 5–26. http://dx.doi.org/10.1177/0013164412449831

Rodríguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98,* 223–237. http://dx.doi.org/10.1080/00223891.2015.1089249

Rodríguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21,* 137–150. http://dx.doi.org/10.1037/met0000045

Rudick, M. M., Yam, W. H., & Simms, L. J. (2013). Comparing countdown- and IRT-based approaches to computerized adaptive personality testing. *Psychological Assessment, 25,* 769–779. http://dx.doi.org/10.1037/a0032541

Salgado, J. F., Moscoso, S., Sanchez, J. I., Alonso, P., Choragwicka, B., & Berges, A. (2015). Validity of the five-factor model and their facets: The impact of performance measure and facet residualization on the bandwidth-fidelity dilemma. *European Journal of Work and Organizational Psychology, 24,* 325–349. http://dx.doi.org/10.1080/1359432X.2014.903241

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331–354. http://dx.doi.org/10.1007/BF02294343

Seo, D. G., & Weiss, D. J. (2015). Best design for multidimensional computerized adaptive testing with the bifactor model. *Educational and Psychological Measurement, 75,* 954–978. http://dx.doi.org/10.1177/0013164415575147

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120. http://dx.doi.org/10.1007/s11336-008-9101-0

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment, 17,* 28–43. http://dx.doi.org/10.1037/1040-3590.17.1.28

Simms, L. J., Prisciandaro, J. J., Krueger, R. F., & Goldberg, D. P. (2012). The structure of depression, anxiety and somatic symptoms in primary care. *Psychological Medicine, 42,* 15–28. http://dx.doi.org/10.1017/S0033291711000985

Sinharay, S. (2013). A note on assessing the added value of subscores. *Educational Measurement: Issues and Practice, 32,* 38–42. http://dx.doi.org/10.1111/emip.12021

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30,* 29–40. http://dx.doi.org/10.1111/j.1745-3992.2011.00208.x

Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 183–206). New York, NY: Routledge/Taylor & Francis Group.

Sunderland, M., Batterham, P., Carragher, N., Calear, A., & Slade, T. (2017). Developing and validating a computerized adaptive test to measure broad and specific factors of internalizing in a community sample. *Assessment*. Advance online publication. http://dx.doi.org/10.1177/1073191117707817

Ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69,* 613–625. http://dx.doi.org/10.1007/BF02289858

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16,* 1–9. Retrieved from http://pareonline.net/getvn.asp?v=16&n=1

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44,* 315–327. http://dx.doi.org/10.1016/j.jrp.2010.03.003

Weiss, D. J., & Gibbons, R. D. (2007). *Computerized adaptive testing with the bifactor model*. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC conference on computerized adaptive testing*. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Zheng, Y., Chang, C. H., & Chang, H. H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 22,* 491–499. http://dx.doi.org/10.1007/s11136-012-0179-6

Received February 23, 2018
Revision received May 21, 2018
Accepted May 30, 2018 ■