Psychometric Methods for Controlling Social Desirability Response Bias in Aggression Questionnaires.

DOCTORAL DISSERTATION

Department of Psychology

2012

UNIVERSITAT ROVIRA I VIRGILI

Cristina Anguiano-Carrasco

Directed by:

Andreu Vigil Colet

Pere Joan Ferrando Piera

A Reo

A shivas

"Son mis amigos los héroes de toda una vida,

dulce emoción que transpone la cruel realidad."

Miguel Abuelo

Acknowledgments

The present work would never have come into being without the inestimable help and support of many people, mentors, friends and family. I would like to acknowledge all those at the Rempeers at UMass, who took me in on a couple of occasions and treated me as one of the group. We shared learning, fun and travel, among other things; Jenna, Kattrina, Amanda, Jerome, Kim, Minji, Sylvia, Rob and Chris, thanks for your kind welcome. I should make special mention of the faculty members Dr. Sireci, Dr. Hambleton, Dr. Keller, Dr. Randall and Dr. Wells, all of whom accepted me in their classes, kindly shared their knowledge and experience, and even took time off from their busy agendas to meet me and talk about my dissertation. I learned a lot in my time at UMass, and I would like to thank everyone there for that. I cannot talk about the nice people at UMass without mentioning Peg Louraine, who I thank very much for all her help. It was much easier to stay at UMass with their help and advice, thanks for everything and for always having a smile for me.

Jenna Copella, I must thank you in particular. You received me in your home and treated me as one of your family. I will never be able to pay you back for all you did for me, especially during November 2012; I will always be grateful for that, so please come and visit me in Spain! My home is your home!

As you may know, it is difficult to express a feeling such as gratitude, in a language that is not your own, particularly if it is not the language you usually use to communicate with the people you wish to thank. So please allow me to skip the protocol and write some words in different languages.

Spanish.

Debo agradecer al Dr. Muñiz, sus palabras de ánimos durante todos estos años de estudiante y su amabilidad al ponerme en contacto con los profesionales de la UMass, facilitándome el poder haber pasado ahí los meses de mi estancia internacional. También agradezco al Dr. Ponsoda su interés por este trabajo y su apoyo durante todos estos años de elaboración.

Sonia Marchal Peñas y Eloi Navarro Serrano, que deciros que no sepáis ya… Gracias por aguantarme, por alentarme a seguir cuando yo hubiera tirado la toalla, por escucharme, ¡por estar ahí! Sabéis lo importantes que sois para mí y no hay palabras suficientes para expresarlo por escrito, pero sí es cierto que sin vosotros cerca, esta tesis no hubiera visto la luz.

A todos mis chicos del Phonesat, en especial Pol Pares, Eduardo Moreno, Manuel Vallecillos y Marcos Adán. Gracias por permitirme formar parte del equipo, por darme esas horas de deshago recorriendo los 100 metros gradas, por permitirme soltar todo ese estrés acumulado. En las derrotas y en las victorias, ¡serios y a por el partido! Gracias también al Sr. Pepe por las conversaciones pre y post partido y a Imma Juncosa, que pese a ser "el enemigo", me acogió en la gran familia del FutSal de la Salle.

Mercedes Blanco Antón, gracias por tu incondicional apoyo en todos mis planes y "locuras"; des de Madrid, Londres, Tarragona, Zamora, Miami, Barcelona... estés donde estés, siempre te siento cerca y tu apoyo ha sido siempre un gran pilar para mí.

También mi más sincero agradecimiento a Héctor Martínez Domínguez, por las horas dedicadas al diseño y la ilustración del formato final de la tesis. Gracias por embellecer mi trabajo.

A toda la gente del departamento de psicología, Catedráticos, Titulares, Agregados, Lectores, Asociados, Ayudantes y Becarios que han compartido conmigo largas horas de trabajo e innumerables conversaciones. Gracias por todo vuestro apoyo, por contribuir con interesantes reflexiones a la elaboración de este trabajo y por acogerme en el departamento. En especial Daniel Rivera y Fàbia Morales, por confortar con su presencia las largas tardes de trabajo en el laboratorio y por compartir esos momentos de descanso con conversaciones de lo más variadas.

Ivethe Espinoza, tu sonrisa y energía positiva siempre son una alegría y junto con Carolina Mayor, ¡cuántas risas nos hemos echado aliviando el estrés y planificando el futuro! Gracias por compartirlo conmigo. Pilar Bonasa, ya sabes, las de método siempre pidiendo lo mismo: "Dame muestra por favor…" gracias por estar siempre dispuesta a colaborar.

Catalan.

A la meva mare Pilar Carrasco i al meu pare Santiago Anguiano. Gràcies primer de tot per donar-me la vida, per crear un ambient familiar afectuós i segur, en el que he crescut i que ha contribuït de manera decisiva que avui en dia sigui qui sóc. Pel gran esforç d'acompanyar-me en totes les meves extraescolars (que no van ser poques!) i per fer l'esforç econòmic que això ha implicat (tan sols ara sóc capaç d'adonar-me del que això ha suposat!) això sí, com a mínim, les classes d'anglès han estat ben aprofitades! Gràcies per estar sempre al meu costat i donar-me el vostre suport! I gràcies també a la meva iaia Carolina Fernàndez, sempre has cregut en mi i m'has donat ànims! Aviat tindré més temps per a donar-te el "achuchón" que encara et dec! Tot i no estar present tot el que voldria, sempre tinc les teves frases de

suport ben presents! Al padrí Carrasco i la padrina Lolín, moltes gràcies per confiar en mi, pel vostre suport i el vostre afecte, descanseu en pau.

Gràcies a les pipotes Maite i Txell, pels caps de setmana de balneari i safareig que tan bé senten un cop (o dos!) a l'any per oblidar la feina i l'estrès; encara que després ens passem el dia parlant de feina.... Per ajudar-me a tenir sempre nous objectius i posar-los per escrit per tal de comprovar l'evolució, per les nits escrivint el diari de pipotes, i xerrant de tot... Gràcies per ser les meves amigues des de que teníem dos anys! Jordi Vila i Dani Haro, la vostra sinceritat en totes les converses fins i tot en allò que no vull sentir sempre han estat de gran ajuda.

A tota la gent de la secretaria de la fcep, en especial a la Leo, sempre eficient i disposada a ajudar, gràcies per facilitar-me els tràmits acadèmics i administratius! I també agraïments a tota la secretaria de suport al deganat de la fcep, en especial a la Rosa Garcia i la Carmen Castillo, amb les que vaig treballar durant quasi dos anys, gràcies per donar-me l'empenteta que hem faltava per poder obtenir la primera beca de recerca. El vostre suport es inestimable i va ser l'inici del que finalitza avui amb aquesta tesi.

Es necessari fer menció especial en aquest apartat a la secretaria del departament de psicologia. Esther, Raquel, gràcies per facilitar la paperassa i fer-ho sempre amb un somriure. Joan, sense la teva ajuda per tramitar el depòsit de la tesis, quan jo no podia ser físicament a Tarragona, tot això no hagués estat possible; per això i per molt més, moltes gràcies! Antoni Masip, gràcies per la teva ajuda en les qüestions tècniques i per compartir coneixements generals i converses en les hores de dinar, cafès, tardes interminables…. Gràcies! Sandra Cosi, sempre present i Carmen Hernández, gran suport i amiga, companya, consellera i mentora. Moltes gràcies a les dues!

A totes les companyes i tots els companys becaris (o no) amb els que he tingut el plaer de coincidir en aquesta etapa. A la gent del màster de "death and destruction", Elisa, Carol, Dolors, Lluís, Núria, Íngrid, Edith, Joan Manel, Mireia; entre tots vàrem aconseguir que el màster intensiu no acabés amb nosaltres!

A l'àrea de metodologia, formada per grans professionals que hem van acollir des del primer moment. Al Dr. Lorenzo-Seva, per donar-me la primera oportunitat a l'àrea com a supervisor de la meva primera beca i per estar sempre disposat a resoldre els dubtes de la becaria, encara que no sigui formalment la seva responsabilitat, Gràcies! Als Drs. Vigil-Colet i Ferrando, per la vostra paciència amb els meus atacs d'histèria, pel vostre suport i per compartir els coneixements de manera tan cordial i propera, més enllà de les obligacions dels directors de tesi. Andreu, espero que les becaries "que t'ha enviat al diable" no hagin acabat amb les teves ganes de seguir format a joves investigadors/es; els teus consells son de valor incalculable i aquesta habilitat per "predir" el que vindrà que et caracteritza no té preu. Em quedo, entre moltes altres coses, amb el refrany castellà que hem vas ensenyar: "Del agua mansa líbreme dios, que de la brava me libraré yo". Pere Joan, han estat moltes les hores que hem compartit parlant de recerca, de docència, de còmics i de Conans! Ha estat un plaer poder aprendre de tu, tot i que seguir el teu ritme de pensament supera a qualsevol! Gràcies per frenar el ritme quan t'ho he demanat i per posar-me les piles quan ho he necessitat. A tots dos, codirectors de la tesi que a continuació es presenta, moltes gràcies!

Finalment, em resta agrair a la persona que més hores ha compartit amb mi, que més m'ha acompanyat en aquest viatge, la que més m'ha aguantat, **Mireia Ruiz Pàmies**: que faria jo sense tu! Treballem juntes des del primer curs de carrera, tan en les tasques acadèmiques com en les

extraacadèmiques: treballs de grup, hores d'estudi a la biblioteca, classes de la carrera, més hores d'estudi a la biblioteca, treball al videoclub, més hores de biblioteca, classes de màster, més hores de biblioteca, recollida de mostra i tots aquests anys de doctorat... Per a mi, sempre ha estat molt reconfortant saber que estaves a prop (a la taula del cantó!), una persona amable i sincera i per damunt de tot tranquil·la! La meva antítesi, moltes vegades, que tan sovint m'ha ajudat a veure les coses des d'un altre punt de vista, a tranquil·litzar-me, a respirar profundament i comptar fins a 10! No oblidaré mai la teva ajuda i companyia quan, en ple període d'exàmens, va morir el meu avi. Si no fos per tu, segurament no m'hagués presentat als exàmens que quedaven, tu hem vas ajudar a repassar el que ja havíem estudiat i em vas donar "palitos de la felicitat", aquelles xuxes del videoclub que tan m'agradaven, i em vas fer somriure! Son tantes les coses que t'he d'agrair que mencionar tan sols una anècdota hem sembla poc. Tampoc podré oblidar mai les celebracions post-exàmens amb la Mònica Fernández i la Marta Balaguer.... "¡No hace falta decir más nada!" I podria continuar indefinidament, explicant des dels milers de cops que m'has deixat els apunts fins a la teva inestimable ajuda per depositar la tesi; però em sembla que si ho expliqués tot, podria redactar un llibre sencer! Crec que ja saps lo molt que valoro la teva amistat i vull acabar la secció d'agraïments amb tu, que has estat sempre i sempre seràs la meva companya i amiga. Moltes gràcies Mirehiya!

Thanks to all of you!

¡Gracias a todos!

Gràcies a tots!

INDEX

0. Prologue …..……I

1. Introduction …….……...……1

1.1. Impact of faking on personality questionnaires. What we know …….…..4

1.2. Detecting faking: Existing methods and recent proposals ..7

1.2.1. Social Desirability scales …….……7

1.2.2. Item Response Theory Approaches …….……10

1.2.2.1. Initial studies …….……..…..…11

1.2.2.2. Research based on Practical Person fit indices ……...13

1.2.2.2.1. Standardized Log-Likelihood Index …….13

1.2.2.2.2. Z3 and F2 indices …….……14

1.2.2.3. Research based on optimal person-fit indices: Structural model-based optimal person-fit procedure ……..16

1.2.3. Hybrid Rasch-Latent Class Modeling ……....18

1.3. Correcting faking: Existing methods and recent proposals ……...……19

1.3.1. Removal of cases ….……...22

1.3.2. Scores adjustments ….……...…..…24

1.3.3. General Factor-Analytic Procedure …....26

1.4. The role of individual differences ……..……29

1.5. Some Final Thoughts about Faking …….……31

1.6. Aggressive behavior. Why we choose it ….……..33

1.7. Objectives and hypotheses …….…….. 37

2. Method …….……... 39

2.1. Participants and Measures: Study 1 …….…… 42

2.2. Participants and Measures: Study 2 …….…….43

2.3. Participants and Measures: Study 3 ……...…...……44

3. Results …….…….45

3.1. Assessing indirect aggression in aggressors and targets: Spanish adaptation of the Indirect Aggression Scales ……...46

3.2. A Structural Equation Model at the Individual and Group Level for Assessing Faking-Related Change ……...53

3.3. Controlling social desirability may attenuate faking effects: a study with aggression measures ……73

4. Discussion ……...97

4.1. Conclusions ……108

5. References ……111

0. Prologue

During my third year as an undergraduate student I realized that research was my interest. From that moment on I worked hard and was awarded a student grant for a research project (2005SGR-00017) in collaboration with the methodology department under the supervision of Dr. Lorenzo-Seva. It was while I was involved in this project that I learned how to use some of the basic methodological tools and had my first contact with faking-related research.

After my graduation as a psychologist, I was asked to join the project "Development of psychometric instruments in the assessment of direct and indirect aggression: A situational approach. (Desarrollo de instrumentos psicométricos en la evaluación de la agresividad directa e indirecta: Una aproximación situacional), for which I was awarded a national grant (BES-2009-014251).

All the above led naturally to the present dissertation: Psychometric Methods for Controlling Social Desirability Response Bias in Aggression Questionnaires. Thus, I managed to link my first contact with research – on social desirability response bias – with the new project in which I was involved, the study of aggression.

I must admit I was lucky that the methodology group had an expert on aggressive behavior and aggression measures, Dr. Vigil-Colet, who has supervised the present dissertation, and an expert on faking and social desirability, Dr. Ferrando, who has acted as cosupervisor.

1. Introduction

"Essentially, all models are wrong, but some are useful"

George E. P. Box

1. Introduction

One of the most important barriers that personality test practitioners have to deal with is faking. Psychological tests are most commonly used for selecting and diagnosing (Zigler, MacCann, & Roberts, 2012) and both situations can prompt test takers to fake. When they are used for selection, test takers are often interested in obtaining a job, and they answer under pressure to give the best possible image or the image that they think the employer wants. When they are used for diagnosis, test takers may fake if the assessment is going to be used to influence, for instance, a decision to give the test taker an economic compensation, early retirement, parole or prison probation, children's custody or asylum probation.

Nowadays, faking is used as a synonym for such terms as response bias, response sets, response styles, response distortion, socially desirable responding or malingering among several others (Hopwood, Morey, Rogers & Ewell, 2007; Hough, Eaton, Dunnette, Kamp & McCloy, 1990; Jackson & Messick, 1958; Paulhus, 2002; Ziegler & Buehner, 2009). These terms have their own definitions, but they all have something in common: there is a source of variation, which is not

the attribute of interest, that systematically affects test scores. Another common characteristic of definitions of faking is that the response distortion aims to give a self-description that helps to achieve an objective or goal.

Paulhus (2002) provides the most commonly used conceptualization of social desirability. He defines it as “the tendency to give overly positive self-descriptions” (p. 50). This definition implies that predictions about someone’s future behavior cannot be made on the basis of a personality test, as the results cannot be trusted.

In this regard, Ziegler and colleagues drew up a person × situation conceptualization of faking. That is to say, the pressure of a particular situation may encourage faking but, depending on his/her personality and values, an individual may or may not fake the test (eg. Ziegler & Buehner, 2009; Ziegler, Toomela, & Buehner, 2009). So faking can be conceptualized as the interaction between the demands of a particular situation and a person’s characteristics.

1.1 Impact of faking on personality questionnaires. What we know.

Faking might have an important impact on personality measures. As stated above, we cannot trust people's scores if faking occurs, because it affects rank order and validity.

Some meta-analyses have shown that faking increases mean scores in both laboratory studies (Viswesvaran & Ones, 1999) and studies on real job applicants (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006). Viswesvaran and Ones (1999) reported an increase in the mean of all five major personality traits, while in real job applicants the magnitude of the increases has been seen to depend on the personality traits, the effect size varying from d = 0.11 on Extraversion to d = 0.45 on Conscientiousness. The faking effects also depend on how relevant the applicant thinks the trait is for the job he/she is currently applying for (Ziegler, Toomela, & Beuhner, 2009).

As far as validity is concerned, it is important to take into account that, when faking occurs, the questionnaire does not measure only the individual differences in the trait of interest but also the individual differences in faking behavior. Therefore, as far as test interpretation

is concerned, validity decreases (Ziegler, MacCann, & Roberts, 2012). Validity, however , is usually measured by the correlation between two tests that measure the same trait of interest or by the correlations that the questionnaire shows with strongly related traits (convergent validity) or with unrelated ones (discriminant validity). If faking occurs, correlations may increase as there is a source of variation that is common to both measures. In fact, evidence found by various studies support this statement: for example, an increased correlation has been found between the Big Five dimensions (Pauls & Crost, 2005; Schmit & Ryan, 1993; Zigler & Buehner, 2009; Ziegler, Toomela & Beuhner, 2009). Ziegler and Buehner (2009) used an experimental design combined with structural equation modeling to establish that correlational increases due to faking can be totally controlled for by modeling a latent variable that captures the individual differences in faking.

On the other hand, some studies have examined the differences in validity between motivated and nonmotivated samples. By using a partially invariant factor-analytic model between job applicants and students, Smith and Ellingson (2002) found that the model-data fit was

acceptable, and showed no difference between the samples. Thus, Zigler, MacCan and Roberts (2012) state that "The issue of how strongly faking might affect the validity of scales remains a particularly fertile area for contemporary research" (p. 11).

In order to assess the rank order issue, Muller-Hanson, Heggestad and Thorton (2003) asked 444 participants to fill in a test. Some of them were in the control group and received standard instructions to provide honest answers, while the others were motivated to fake as they were told that only a few selected participants would move on to the next phase, for which they would be paid. Then all the questionnaires were mixed together and the high performers selected. It was concluded that the fewer participants are selected, the fewer control group participants are chosen. Consequently, faking potentially has an important effect on rank order and selection decisions.

1.2. Detecting faking: Existing methods and recent proposals

1.2.1. Social desirability scales

Social desirability scales were the first attempt to control response bias. It is the most commonly used method to operationalize faking (Kuncel, Borneman, & Kiger, 2012). They are used either as a proxy or a direct measure of faking behavior. Social desirability scales consist of items that refer to (a) behaviors that society considers to be good and desirable but which are highly improbable or (b) undesirable behaviors that occur frequently. One of the most famous scales is the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960). Social desirability scales can be used either as a separate instrument or embedded within a general personality questionnaire, as is the case for the Lie scale in the EPQ-R by Eysenck and Eysenck (1975) among others.

The main idea of using a social desirability measure as a faking detection instrument is that if someone is trying to give an unrealistically good image of themselves, their scores will be very high

because they will have endorsed most of the desirable items which, as mentioned above, are highly unlikely to happen in real life. Although research into these instruments sounds promising, the results have been disappointing. On the positive side, if subjects are instructed to fake in laboratory research, the social desirability scales do quite a good job of differentiating the fakers from the honest respondents (Viswesveran & Ones, 1999). On the negative side, however, the scales do not correlate with job performance and their use as a faking control instrument does not improve predictive validity (Ones, Viswesvaran & Reiss, 1996).

If social desirability scales are used to remove the highly scoring respondents, well-behaved respondents may also be removed because they are mixed in with fakers. In this regard, when a social desirability scale is administered using standard instructions and under neutral conditions (i.e. no pressure for faking), there seems to be agreement that its scores essentially measure a personality trait (Ferrando, Chico, & Lorenzo-Seva, 1997; Katz & Francis, 1991; Loo, 1995; Lajunen & Scherler, 1999). Interpretation of the scores under faking-good-motivation conditions (in high-stakes assessment or under

appropriate instructions) is more complex. Eysenck and Eysenck (1975) hypothesized that, in this case, the social desirability scale behaves as it should and serves to detect dissimulation. This double interpretation is quite general, and allows several hypotheses to be considered. The most complex scenario is that the scale measures a different factor (or perhaps more than one) with different item measurement properties (Michaelis & Eysenck, 1971). Conceptually this means that, under faking-motivation conditions, respondents attach a different meaning to the items. However, this double interpretation of social desirability scales is still controversial. Research carried out by Ferrando and Anguiano-Carrasco (2011a) examined the prediction power that social desirability scores have, when respondents are allowed to respond under neutral conditions, on the score increments due to faking-inducing instructions. Their results showed that neutral social desirability scores do not correlate with scale increments caused by faking and supported the hypothesis that under neutral conditions social desirability scores measure something other than the propensity to fake.

The unlikely virtues scales, in which subjects are asked about virtues they are unlikely to have, are a similar case in point. It has been shown that these scales can detect intentional distortion when laboratory-induced fakers are compared with honest respondents (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990). But, as happens with social desirability scales, corrections using unlikely virtues scales do not generally improve scores (Christiansen, Goffin, Johnston & Rothstein, 1994; Ones, Viswesvaran & Reiss, 1996).

1.2.2. Item Response Theory Approaches

Item Response Theory (IRT; e.g. Hambleton & Swaminathan, 1989; Muñiz, 1997) approaches are mostly based on fit indices assessed at the individual level, and usually known as person-fit indices. These indices assume that even when a particular IRT model globally fits data, there might still be a percentage of individuals for whom the model is not appropriate. So person-fit indices take into account how the individual data agree with the model, in a similar way to outlier detection techniques. Person-fit indices aim to detect aberrant respondents who, in the case of the present text, can be

identified as fakers. Most person-fit indices used in faking research are parametric indices based on the likelihood function. They work by figuring out the maximum likelihood value of theta (θ – magnitude of the latent trait of the individual) and then examining some version of the likelihood function that can be used to determine if a level of theta is likely or unlikely (Zickar & Sliter, 2012). The main concerns when these approaches are used are the number of hits (number of fakers identified) and false positives (number of honest respondents identified as fakers) raised by the person-fit index classification. The question that needs to be answered, however, is what false-positive rate is an organization willing to tolerate (Zickar & Sliter, 2012).

1.2.2.1 Initial studies

The first study which used person-fit procedures to assess faking appears to be that of Zickar and Drasgow (1996). They based their research on the changing-persons paradigm. More precisely, they used both practical person-fit indices and optimal person-fit indices to test the hypothesis that faking will produce some intra-individual

inconsistency in the response pattern. For the optimal indices they modeled the likelihood ratio test as a shift of +0.50 to the right of the theta scale for those items that were deemed fakeable. No shift occurred in the remaining items. They used the personality scales of the Assessment of Background and Life Events (ABLE- Whithe, Nord, Mael & Young, 1993) and clustered the respondents in three groups: honest respondents, instructed to fake-good respondents and trained to fake-good respondents. Zickar and Drasgow (1996) used dichotomous and polytomous IRT models to fit the data and estimated the person-fit indices in order to examine the extent to which the indices could differentiate trained fakers and instructed fakers from honest respondents. They found that standard person-fit indices were about as effective as social desirability scales at distinguishing fakers from honest respondents. Optimal person-fit indices produced little improvement.

On the basis of this initial study, other research was carried out. We should mention three studies that are particularly relevant: two on practical person-fit indices and one on optimal person-fit indices.

1.2.2.2. Research based on practical person-fit indices

1.2.2.2.1. Standardized Log-Likelihood Index

Ferrando and Chico (2001) used the Extraversion, Neuroticism and Psychoticism scales of the Spanish version of the Eysenck Personality Questionnaire-Revised (Aguilar, Tous & Andrés, 1990) to compare the use of IRT models and social desirability scales (in this case, the Marlowe-Crowne Social Desirability Scale [1960]) for detecting faking. Like Zickar and Drasgow (1996), they based their approach on the changing-persons paradigm, but applied it to well-known "civil" tests. They used only a practical index: the standardized log-likelihood index (*lz* – Drasgow, Levine & Williams, 1985). The *lz* index is based on the idea that the likelihood function of a given response pattern will be larger when the pattern fits the overall IRT model and smaller when it does not, as happens with the general log-likelihood (*lo*) index. The standardized index was chosen because it is robust to the effects of test length and model choice and can be interpreted like a *z* – score. They found that for practical purposes person-fit indices were not

useful for detecting faking, although they performed about 10.2% above the level of chance. The detection accuracy obtained with the Social Desirability scale was 28.6% higher than chance.

1.2.2.2.2. Z3 and F2 index

Z3 is a practical person-fit index based on the likelihood function after controlling for its variability across the latent trait level; more precisely, it is based on the height of the likelihood function. A lower Z3 indicates that response patterns and item characteristics differ. F2 is another practical index, but it is based on the comparison between the participant's given responses and the predicted ones. The predicted responses are calculated using the item characteristics and the respondent's estimated theta level. The higher the value of F2, the higher the discrepancy between the predicted and the real responses and the more likely it is that the responses do not reflect the respondent's true level. In summary, decreased values of Z3 or increased levels of F2 indicate internal inconsistency that will be interpreted as response distortion. The appropriateness of Z3 and F2

indices for detecting fakers has been tested by two main studies. The first one was by Brown and Harvey (2003). Respondents first filled in the Agreeableness and the Conscientiousness scales from Brown's Five-Factor Model (1997) and were then asked to answer honestly, to fake as positively as possible or to fake realistically, thus creating three groups. In no cases were the positive fakers or the realistic fakers appropriately detected whichever appropriateness index was used. The second study was conducted by Harvey, Wilson and Hansen (2005). Their respondents were state police officers who twice completed the Responsibility scale of the California Personality Inventory (Gough & Bradley, 1996). The first time they were asked to answer honestly and the second time they were asked to imagine they were applying for a position as a state trooper and told that faking was an acceptable method of responding. When the two groups were compared they found that the F2 index values were not significantly different, but the Z3 values were. Although it can be considered a success, this statistically significant difference was of no practical use for identifying faking. The results regarding the F2 and Z3 indices were not surprising, as they are less accurate than the *lz* index. If *lz* was not

able to reach an acceptable degree of detection, it is only to be expected that F2 and Z3 would not provide better results.

1.2.2.3. Research based on optimal person-fit indices: Structural model-based optimal person-fit procedure.

This method is based on optimal indices, but differs from that used in Zickar & Drasgow's (1996) study because this new method is based on the changing item paradigm instead of the changing person paradigm. Ferrando and Anguiano-Carrasco (2012) adopt the modern view of the theta-shift mechanism, and consider that, in general, faking is expected to produce some intra-individual inconsistency in the response pattern (i.e. not all the items will be equally affected) but that this inconsistency is so subtle that it cannot generally be detected by standard or practical person-fit indices. The basic idea is to identify whether a particular response pattern better fits an "honest" profile or a "faking" profile. This point has been considered by several authors (Kuncel & Borneman, 2007, Zickar & Drasgow, 1996, Zickar & Sliter, 2012) and is the basis of the procedure proposed by Ferrando and

Anguiano-Carrasco. Their proposal is a two-stage procedure. In the first stage a partially invariant item factor-analytic model is fitted simultaneously in the two-groups or waves. Then, if the fit is considered acceptable, the item parameter estimates are: (a) reparameterized so that they are transformed into the parameters of the IRT model, and (b) taken as fixed and known and used as input for the second stage. Overall, the aim of the first stage is to obtain a model-based 'typical' inconsistent pattern as an alternative hypothesis to consistency. Because this pattern is model-based the procedure is expected to be more powerful than the initial proposal by Zickar and Drasgow (1996) because in their study the alternative inconsistent pattern was determined ad-hoc by using a constant shift on some items.

In the second stage, the two sets of calibrated item parameters are used to compute a likelihood-ratio based optimal person-fit index. The outcome of the ratio is expected (to some extent) to identify respondents who faked the measurement instrument. The second stage is the same as in Zickar and Drasgow's original study (1996). However, because the alternative pattern is better specified in the

present procedure the second stage likelihood-ratio test is expected to be more effective.

The participants were instructed to respond honestly the first time and to fake the second time. All respondents filled in the Extraversion, the Neuroticism and the Psychoticism scales of the EPQ-R Spanish version (Aguilar, Tous & Andrés, 1990). The authors obtained a hit rate of 66% with a false alarm rate of 5%, which is, as far as I know, the best hit/false-alarm rate obtained so far.

1.2.3. Hybrid Rasch-Latent Class Modeling

Hybrid Rasch-Latent Class Modeling combines the benefits of traditional Rasch modeling and latent class modeling, and also enables subgroups to be identified a posteriori using response patterns generated by individuals in each group. Holden and Book (2009) asked participants to imagine being in a selection process for the military and answer the NEO-Five Factor Inventory (NEO-FFI – Costa & McCrae, 1992) and a set of social desirability items from the Balanced Inventory of Desirable Responding (BIDR – Paulhus, 1998). The

respondents were grouped according to the three types of instructions received: honest responding, fake-good responding (in order to be accepted) or fake-bad responding (in order to be rejected). They tested the potential of the Hybrid Rasch-Latent Class Modeling technique for detecting the three response patterns and also for determining whether validity increased. The technique effectively differentiates between the response patterns of each group, but despite the promising results it should only be used in real selection processes with caution. They also found that validity was better than when only the social desirability measure was used. Hybrid Rasch-Latent Class Modeling, then, is the first technique to have encouraged further efforts to be made in detecting faking in personality measures.

1.3. Correcting faking: existing methods and recent proposals

The idea underlying all the correction methods described above below is the same. The correcting techniques use social desirability scales or items (depending on the method) to correct content scores. It has been proved that when faking occurs social desirability scores increase

(Viswesveran & Ones, 1999) and the mean scores on the trait of interest scales will also increase if the trait is desirable, or decrease if it is undesirable. So, if we correct for social desirability, the trait scores will also be affected (decreasing in the case of desirable traits or increasing in the case of undesirable ones). Thus, correcting for social desirability is also supposed to correct for faking or, at least, to mitigate the effects that faking has on the trait scores. That is, when we correct for social desirability a "positive side effect" is that faking effects might be reduced. Figure 1 illustrates the idea that underlies the use of social desirability control methods to correct for faking.

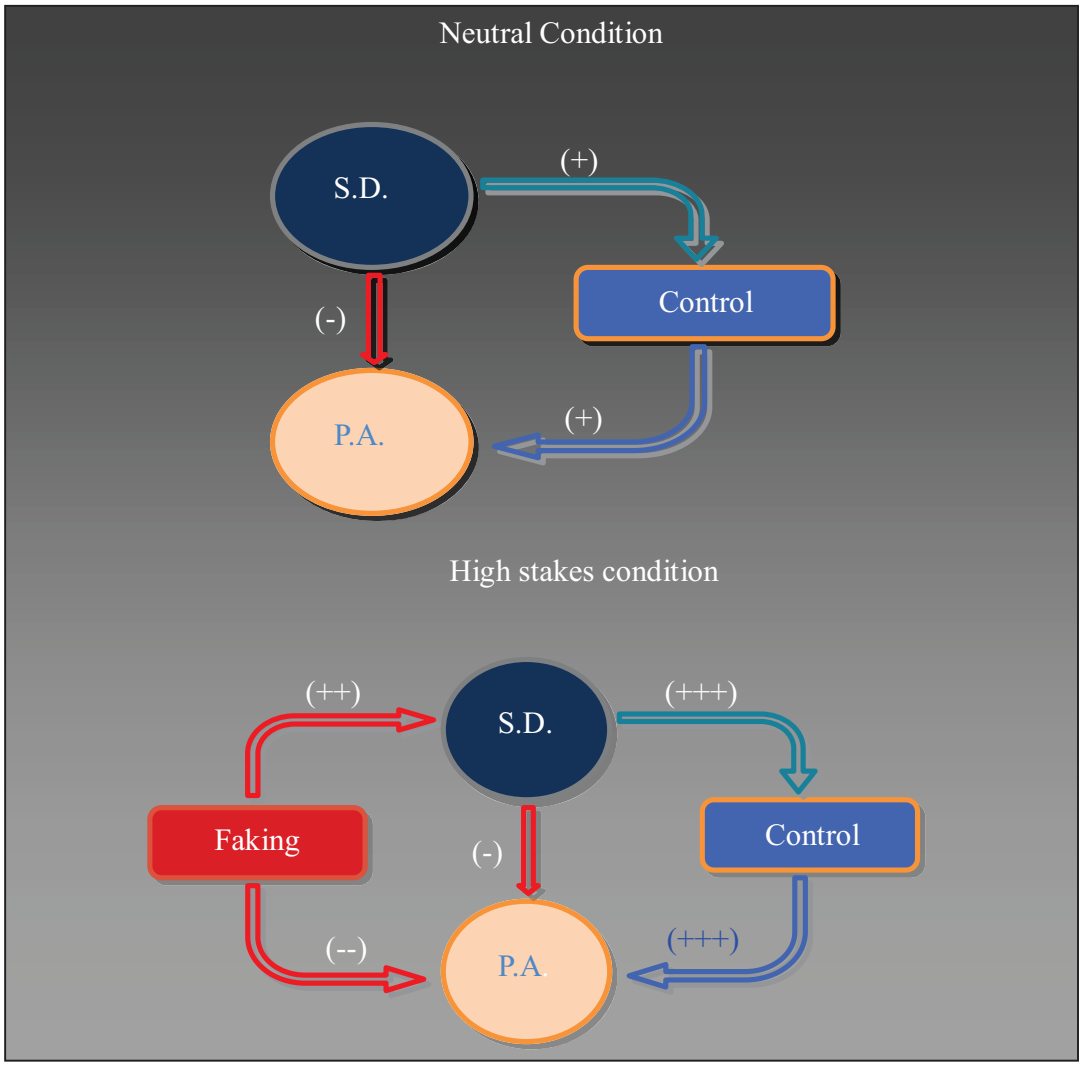


Figure 1. Idea underlying the use of social desirability control methods to correct for faking

1.3.1. Removal of cases

This technique consists of deleting or removing either the respondents who scored high on a social desirability scale or the respondents detected as fakers no matter what detection technique is used. The removal of cases has been the subject of a considerable amount of research but results have been inconsistent. Some studies show that it can be positive (Pannone, 1984; White, Young, Hunter, & Rumsey, 2008; among others) while others show no increase in validity (Christiansen, Robie, & Bly, 2005; Dudley, McFarland, Goodman, Hunt, & Sydell, 2005; Hough, 1998; Hough, Eaton, Dunnette, Kamp, McCloy, 1990). Christiansen, Robie & Bly (2005) examined how faking could impact criterion-related validity by comparing respondents identified as fakers with respondents identified as honest. They found no significant differences between the two groups although the faking respondent's criterion-related validity seemed to be slightly lower. In a simulation study, Schmitt and Oswald (2006) examined how removing applicants on the basis of their faking scores affects criterion-related validity and job performance. They found that the procedure had a negligible effect on both criterion validity and job performance. Using

an Unlikely Virtues scale to remove applicants, Hough (1998) found variable results in three different samples. He found that 90% of the decisions match with the decisions made on the basis of observed scores in the first sample, 69% of the decisions in the second sample and 54% in the third. Thus, the results of removing cases to correct for faking have been inconsistent, and there is no consensus on what the appropriate cut-off value is for flagging applicants as fakers. Furthermore, case removal has two major problems. Firstly, extreme-score respondents may be nice, well-behaved people who are wrongly identified as fakers and so wrongly removed (Smith & Ellingson, 2002). Secondly, if social desirability is related to the trait of interest, removing extreme-score respondents may mean removing respondents who have extreme scores on the trait we are attempting to measure (Vigil-Colet, Morales-Vives, Lorenzo-Seva, Camps & Tous, in press; Jackson, 1989). Thirdly, removing extreme individuals does not give the psychologist non-faked scores, so faking still affects the remaining individuals' scores. Taking into account these limitations and that neither criterion-related validity nor job-performance prediction are improved, this technique could be considered inappropriate.

1.3.2. Score adjustments

Several studies have assessed the outcomes associated with adjusting scores mainly by using detection scales such as social desirability scales. Most of the studies that use score adjustment techniques have shown that criterion-related validity is negligibly impacted by these score corrections (e.g., Borkenau & Ostendorf, 1992; Christiansen, Goffin, Johnston & Rothstein, 1994; Dudley, McFarland, Goodman, Hunt & Sydell, 2005). The traditional approach to score adjustments is to partial out social desirability scale scores from the relation between content scores and performance. In this regard, two meta-analyses have shown that criterion-related validities decrease slightly when social desirability is partialled out. The most relevant issue studied from this point of view is the impact that corrections may have on rank order and, therefore, on selection or hiring decisions. Christiansen, Goffin, Johnston & Rothstein (1994) showed that 85% of respondents changed their rank order when corrections were made. Rosse, Stecher, Miller & Levin (1998) found similar results. However, Ellingson, Sackett & Hough (1999) suggest that score adjustments

appear to have little impact on the proportion of correct selection decisions.

Although the rationale for this technique is quite logical, some limitations should be mentioned. First, as mentioned above, criterion-related validity is hardly affected when this technique is used, so there is no big improvement in observed scores. Second, all items should be considered as parallel measurements of the latent trait if the techniques is to be used appropriately and this is rarely the case (Leite & Cooper, 2010). Third, it should also be noted that partialling depends on the social desirability test the practitioner uses. As there is no unitary definition of social desirability, different scales may be measuring different concepts (Paulhus, 1991) so when the social desirability scale scores are partialled out, exactly what is removed will depend on the scale. Fourth, a second test is needed, so administration time increases and the probability of boredom or fatigue effects will also increase. Fifth, from an applied point of view, the practitioner cannot easily obtain individual trait estimators free of social desirability bias, so the free–of-bias results are only based on group estimations. Some authors have proposed that subjects' scores

can be modified by adding or subtracting 0.5 points per item depending on the item fakability (Zickar & Drasgow, 1996). This may be considered an approximation, but it would not give subjects' free-of-bias scores. Sixth, social desirability is regarded as a personality domain that is related to other personality traits that may be of interest (e.g. agreeableness), so if social desirability is removed from the correlations, some content validity of the trait of interest is also removed (McCrae & Costa, 1983; Vigil-Colet, Morales-Vives, Lorenzo-Seva, Camps & Tous, in press).

1.3.3. General factor-analytic procedure

Recently, Ferrando, Lorenzo-Seva and Chico (2009) proposed a general factor-analytic procedure for assessing response bias in questionnaire measures. The procedure has two main steps. The first step identifies a factor related to social desirability. To do so, a set of items related to social desirability is selected. These items are known as markers. The inter-marker correlation matrix obtained is analyzed using factor analysis and the corresponding loading values of each marker on the

social desirability factor are taken as fixed and known. These loading values are then used with the Instrumental Variables Technique (Hägglund, 1982) to compute the loading values of the content items on the social desirability factor, and the variance explained by the social desirability factor is removed from the inter-item correlation matrix. In the second step, the residual inter-item correlation matrix is analyzed using factor analysis to identify the content factor or factors of interest that are orthogonal to the social desirability factor. The application of this procedure at the item calibration level provides two loading estimates for each item: a loading on the content factor that the test wants to measure, and a loading on an orthogonal factor identified as social desirability. Thus, social desirability-free content scores are obtained and the content validity for the personality factors remained the same. Vigil-Colet, Ruiz-Pàmies, Anguiano-Carrasco and Lorenzo-Seva (2012) used this procedure to assess the impact that social desirability has on aggression questionnaires. They found that there were no differences in the correlations between different scales that measure the same trait (aggressive behavior in this case) whether social desirability was included or removed. This indicates that the

procedure does not give a residual matrix of spurious correlations, and that criterion validity is unaffected.

The general factor-analytic procedure overcomes most of the problems associated with adjustment methods explained above. First, all items are independently calibrated so there is no need to assume that they are parallel measures of the latent trait. Second, the procedure uses only a few markers, which are selected because of their high loadings on a social desirability factor, so it depends not on a particular scale but on the psychometric properties of the selected markers. Third, the practitioner only has to add a few items (Ferrando, 2005, recommended that four or five items may be enough) to the content test he/she is interested in. Thus, administration time is almost the same and such confounding variables as fatigue or boredom are prevented. Fourth, the procedure enables individual free-of-bias scores to be obtained, which is by no means easy with other methods, and not only group scores. Fifth, as the procedure uses an orthogonal rotation method, it is assumed that no content validity will be removed from the trait of interest. In fact, recent research using a test based on the five factor models developed with this

methodology has shown that the convergent validity of these personality dimensions remained almost unaffected by the procedure (Vigil-Colet, Morales-Vives, Lorenzo-Seva, Camps & Tous, in press).

To summarize, the general factor-analytic procedure seems to override almost all the limitations that adjustment methods have. This is the main reason for choosing this procedure in the present study.

However, the recent study by Ferrando and Anguiano-Carrasco (2011a) mentioned above raises some doubts as to whether social desirability scales or items are appropriate for eliminating faking so this hypothesis should be tested.

1.4. The role of individual differences in faking.

One of the least studied issues in the field of faking is the role of individual differences. As recent research has found (Burns & Christiansen, 2006; Mersmer-Magnus & Viswevaran, 2006; Rothstein & Goffin, 2006) one of the basic issues that still needs to be resolved is how individual differences affect the amount of change estimated under similar pressure conditions. Several studies consider faking to

be an individual differences variable (Furnham, 1986; Lautenschlager, 1986; McFarland & Ryan, 200,2006; among others) so it is of interest to study if under the same conditions all respondents change their scores in the same direction and magnitude or if it is an individual-differences characteristic. It is also of interest to examine whether different personality traits are affected differently by faking because, as far as we know, this question is still unsolved.

One of the most important aspects of this topic is that faking has an adverse effect on decisions in a selection process (Zigler, MacCann & Roberts, 2012). This effect on the selection process is due to the fact that rank order will depend on whether people fake their answers and how well they fake. Several studies have shown that participants who fake are often selected as their scores are achievement-oriented. Likewise, if only a few participants are to be selected, the faking effect acquires greater importance (eg. Muller-Hanson, Heggestad & Thorston, 2003) as rank order can be strongly affected. Individual differences in faking are expected to increase or decrease this effect as a function of the degree of individual differences in faking on the trait measured.

1.5. Some final thoughts on faking

Research on faking has a long tradition that can be traced back to the beginning of the last century and is still an important research issue nowadays. A revision of the numerous existing methods clearly shows that there is still a lot to do and much to improve. As far as detection methods are concerned, it has been demonstrated that no method works on an acceptable ratio of hits/false-alarms. Although the results obtained by the structural model-based optimal person fit procedure is encouraging, they need to be replicated and polished before they can be generalized to different measures and samples. The use of social desirability scales to detect fakers or correct personality-measure scores is also controversial as they increase test length and there is no evidence to show that they can effectively predict faking (Ferrando & Anguiano-Carrasco, 2011a). As mentioned above, the most popular correction method so far – partialling social desirability out from the measurement-job performance correlation – has many limitations and is thus not reliable method if decisions are to be made. Furthermore, legally it is controversial if candidates are excluded from a selection process on the basis of their scores being detected as

faked, particularly because research on this topic points out that results are inconsistent. The same can be said of correction methods. As no method has proved to improve validity or job-performance prediction it is hard to justify decreasing someone's scores in a diagnostic or selection process.

One issue that still needs to be fully discussed in the faking literature is the impact that individual differences have on the amount of change due to faking. This topic should be of great interest considering the impact it has on rank order and consequently on selection processes. If a particular personality trait is not impacted by individual differences, which means that all respondents change their scores in the same direction and by the same magnitude, faked scores will be different from neutral scores but rank order will not be affected. Therefore, controlling or correcting faking may not be so important for these measures. On the contrary, if a personality trait is highly impacted by individual differences the rank order will be deeply affected, selection decisions may not be the optimal, and the selection of an inappropriate candidate will have economic consequences. It should also be borne in mind that selection processes are not cheap.

Another important limitation that should be pointed out is that all existing correction methods are based on social desirability scales or items. Traditionally, social desirability and faking have been viewed as the same concept although they are not, as several studies have shown (e.g. Ferrando & Anguiano-Carrasco, 2011a). Therefore, we can only be sure that we are correcting for social desirability but not for faking. The present study will also deal with this issue.

1.6. Aggressive behavior: Why we choose it

There are several definitions of aggression, which are not totally equivalent. The most accepted definition is "any behavior that is meant to damage, physically or psychically someone" (Berkowitz, 1996). This definition implies that the behavior that produces damage must be intentional, and not beneficial for the target, and that we should be concerned not only with physical damage but also psychic damage.

Aggression consists of cognitive, emotional and instrumental components. Usually, cognitive components are those associated with

hostility, emotional components with anger and instrumental components with aggression (Berkowitz, 1993). For the purposes of the present study, we will focus on aggression, because it has been found that measuring its emotional and cognitive aspects is extremely complex (Eckhardt & Deffenbacher, 1995; Suarez & Williams, 1989), and that these aspects cannot be considered aggression but the cognitions and emotions associated with it.

Instrumental aggression is defined as the cold-blooded, motor component of aggression, which means any behavior intentionally performed to harm someone (Berkowitz, 1993). Within this instrumental component we can distinguish between indirect and direct aggression. Indirect aggression is a type of aggressive behavior that is not directly manifested against the attacked person. This form of aggression involves some kind of social manipulation in which the aggressor acts on the people around the attacked person with the sole aim of harming him or her without entering into confrontation (Björkqvist, Osterman & Kaukiainen, 1992). This kind of aggression – also known as indirect, social or relational (depending on slight nuances) – emerges during the socialization process of an individual

(Vaillancourt, 2005), considered by some authors as the most common aggressive behavior in adulthood (e.g., Lagerspetz & Björkqvist, 1994). The main characteristics of indirect aggression are the avoidance of face-to-face confrontation and the use of the social environment as a source of damage.

Direct aggression, in contrast, is overtly manifested and the aggressor directly faces the target. Direct aggression is divided into two principal components: physical aggression and verbal aggression. Physical aggression occurs by the direct impact of the body or an instrument on the target (Berkowitz, 1994; Björkqvist, 1994) and includes punching, kicking or pushing. Verbal aggression occurs through the use of language and includes mocking, insulting, taunting or sarcasm (Berkowitz, 1994; Björkqvist, 1994).

Although indirect aggression is considered to be the most socially acceptable form of aggression, as it emerges with individual socialization, in general aggressive behaviors are considered highly undesirable in our society. Therefore, faking is expected to have greater effects on these measures as individuals want to give a good

impression when they fake. This is one of the main reasons aggression measures have been used in the present study.

Some studies have shown that aggression measures can be deeply impacted by social desirability. In general, the research on this issue has shown a moderate-to-high relationship between social desirability and aggression measures. In this respect, Biaggio (1980) and Selby (1984) reported that most of the correlations between the Buss-Durkee Hostility Inventory scales and the Marlowe-Crowne Social Desirability scale were in the range r = -.3 to -.5. Social desirability has also been related to measures of violent behaviors and partner abuse (Bell & Naugle, 2007; Devon, Colley & Walkey, 2004) and those aspects of NEO-PI-R scales most related to aggressive behavior such as impulsivity and angry hostility (Holden & Passey, 2010). Recent research by Vigil-Colet, Ruiz-Pàmies, Anguiano-Carrasco and Lorenzo-Seva (2012) using the general factor-analytic procedure showed that items on aggression questionnaires have moderate-to-high loadings on a social desirability factor and that, when corrected for this effect, aggression scales tend to increase their scores considerably. This is another reason why these measures have been used here.

1.7. Objectives and hypotheses

Reviewing the questionnaires available in Spanish to assess aggressive behaviors, it was surprising to find that none of them assess self-reported indirect aggression in adulthood. Therefore, our first objective was to adapt a valid, reliable self-informed questionnaire. We selected the Indirect Aggression Scale by Forrest, Eatough & Shevlin (2005) as the questionnaire most appropriate for our purposes.

A second objective was to shed some light on the faking literature and provide greater insight into how individual differences affect faking behavior. To achieve this objective, a new procedure for assessing the amount of trait-level change due to faking will be developed. The procedure will allow users to estimate the amount of variance in the change scores due to faking and also to assess the amount of variance that can be explained by individual differences. We also hypothesized that individual differences will have an important impact on the change scores due to faking. We also tested if different personality measures are impacted to the same extent by individual differences. We hypothesized that such personality traits as highly undesirable

measures will be less impacted by individual differences as all subjects will fake in the same direction and by a similar magnitude. However more 'controversial' traits are expected to be impacted to a greater extent.

Finally, we tested the assumption that correcting for social desirability will eliminate or mitigate the impact of faking. We used aggressive behavior measures to test the hypothesis, as they are considered to be highly socially undesirable. Thus, when faking occurs, the scores on aggression questionnaires are expected to decrease in comparison to honest responses. Furthermore, when faking occurs the scores on social desirability scales are expected to increase (see Viswesveran & Ones, 1999). So, we hypothesized that the use of the general factor-analytic procedure will eliminate social desirability from content scores, so faking will also be eliminated or, at least, mitigated.

2. Method

2. Method

The results are explained below in an individual paper focusing on each of the objectives and hypotheses. The method used in each paper is described in detail; therefore, here I will explain the general design used in two of the three papers (Ferrando & Anguiano-Carrasco, 2012; Anguiano-Carrasco, Vigil-Colet & Ferrando, in press), which is somewhat complex.

The study uses a two-wave two-group design. The two groups are the control and the experimental groups differentiated by the instructions they received. The control group participants received standard instructions both times and the experimental group participants received standard instructions the first time and faking-inducing instructions the second. The standard instructions were designed to make the participants answer honestly and avoid overthinking. Personality is not correct or incorrect, it is just what it is. Faking-inducing instructions prompted the participants to imagine themselves in a selection process for a job they are really interested in and therefore to try to give a good impression, to put themselves in a

favorable light, in order to increase their chances of getting the job. The retest interval was 6 weeks because it is the minimum time needed to avoid memory effects (Ferrando, 2002). Figure 2 shows the two-wave two-group design and figure 3 shows the instructions that each group received for each wave.

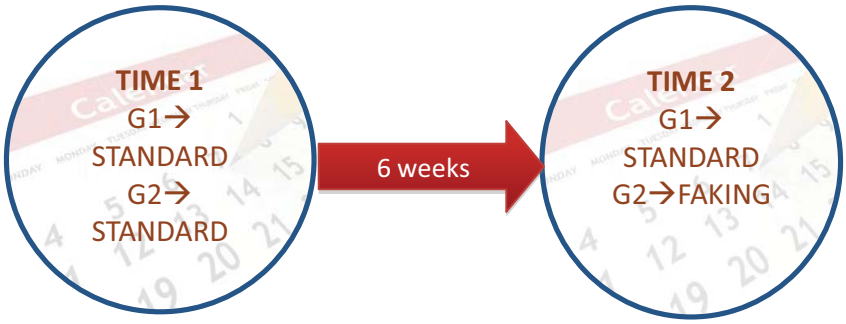


Figure 2. Two-group two-wave design

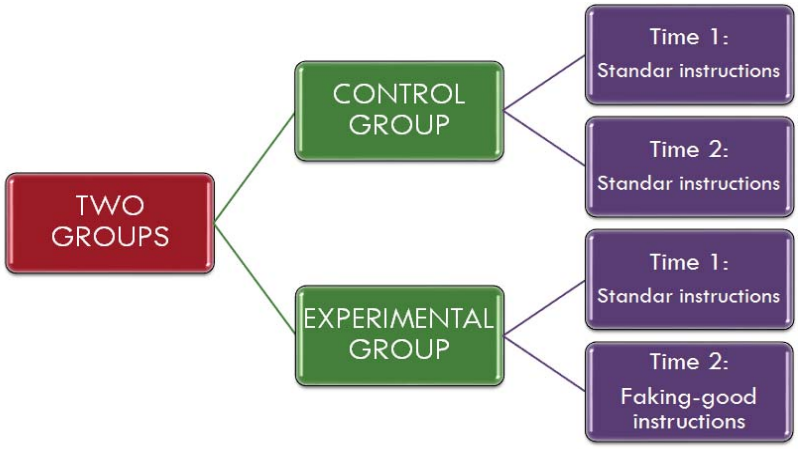


*Figure 3.*Instructions that each group received for each wave

2.1. Participants and measures: Study 1

Participants

Participants were 935 undergraduates aged between 17 and 50 years old (mean age 23.47; s.d. 6.74). A total of 46.42% of the sample were men and 53.58% were women.

Measures

To achieve our first goal, we adapted the Indirect Aggression Scale by Forrest et al. (2005) in its two forms: aggressor and target. To assess validity, the Spanish version of the Dickman Impulsivity Inventory (DII- Chico, Tous, Lorenzo-Seva & Vigil-Colet, 2003) and the Bus and Perry Aggression Questionnaire, Spanish version (BPAQ- Morales-Vives, Codorniu-Raga & Vigil-Colet, 2005) were used.

2.2. Participants and measures: Study 2

Participants

To achieve the second goal, 512 Psychology and Social Science undergraduates participated. They were randomly assigned in classroom groups to the control or experimental group. The control group consisted of 235 students (mean age 19.85, 77% women) and the experimental group of 277 students (mean age 21.54, 75% women). Because the descriptive statistics of both groups were so similar, “quasi-comparability” was assumed.

Measures

All participants twice filled in the Lie scale of the Spanish version of the Eysenck Personality Questionnaire (Eysenck & Eysenck, 1975) under the conditions of the group they had been randomly assigned to (see Figure 3).

2.3. Participants and measures: Study 3

Participants

Participants were 371 university students who filled in the questionnaires twice, following the instructions received in each case (see figure 3). “Quasi-comparability” was assumed for both groups as they were similar in age (mean 21 years old) and they both consisted of approximately 85% women.

Measures

To achieve the third goal, participants completed the Physical Aggression Scale and the Verbal Aggression Scale on BPAQ, Spanish short version (Morales-Vives, Codorniu-Raga & Vigil-Colet, 2005), as well as the Indirect Aggression Scale, Spanish short version (Anguiano-Carrasco, & Vigil-Colet, 2011) created to achieve the first goal.

3. Results

Psicothema 2011. Vol. 23, nº 1, pp. 146-152
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Assessing indirect aggression in aggressors and targets: Spanish adaptation of the Indirect Aggression Scales

Cristina Anguiano-Carrasco and Andreu Vigil-Colet
Universidad Rovira i Virgili

In recent years, there has been increasing interest in indirect aggression as the most common aggressive behaviour in adulthood. Despite this interest, there are not a great many instruments for measuring this behaviour in adults. The aim of our study was to develop the Spanish adaptation of one of the few instruments that does exist: the Indirect Aggression Scale, in its aggressor and target versions. The analysis of these scales in a sample of 935 university students showed that the aggressor and target versions of the scales had good reliabilities, but that a one-factor structure seemed more feasible than the three-factor structure initially proposed. Taking this one-dimensionality, we developed short versions of the scales, which also showed good reliabilities. The aggressor version presented good convergent validity with direct aggression and impulsivity measures. Finally, none of the scales showed differences associated with sex.

Evaluando la agresividad indirecta en agresores y víctimas: adaptación española de las Escalas de Agresividad Indirecta. En los últimos años se ha producido un creciente interés en la agresividad indirecta debido a que es la forma de agresividad más frecuente en la edad adulta. A pesar de ello no existe un gran número de instrumentos de medida para adultos de la misma. El principal objetivo del presente estudio es desarrollar una adaptación en español de uno de los pocos instrumentos disponibles: las Escalas de Agresividad Indirecta en sus versiones de agresor y víctima. El análisis de dichas escalas en una muestra de 935 estudiantes universitarios mostró que ambas formas presentan una buena fiabilidad pero que en ambos casos la estructura factorial de las mismas es unidimensional, en lugar de la estructura de tres factores propuesta por los autores. Teniendo esto en cuenta se plantea la posibilidad de desarrollar una escala reducida de un menor número de ítems. La versión para agresores presentó una buena validez convergente con otras medidas de agresividad y de impulsividad. Finalmente, no se observaron diferencias asociadas al sexo en ninguna de las escalas.

Traditionally the study of aggressive behaviour has focused on direct aggression. However, in recent years, there has been increasing interest in other kinds of aggressive behaviour that are not usually directly manifested against the attacked person. These forms of aggression involve a sort of social manipulation in which the aggressor acts on the people around the attacked person with the sole aim of harming him without having to face him directly (Bjorkqvist, Osterman, & Kaukiainen, 1992). This kind of aggression —also known as indirect, social or relational aggression (depending on slight nuances)— appears during the socialization process of individuals, so that the physical or verbal aggression types typical in children and adolescents turn into other kinds of aggression in adults (Vaillancourt, 2005). So, while physical aggression reaches a peak at around 30 months of age, after which it shows a progressive decrease, indirect aggression begins during childhood and progressively increases until it peaks during adolescence and adulthood (Bjorkqvist et al., 1992; Tremblay, 2005; Cangas, Gázquez, Pérez-Fuentes, Padilla, & Miras, 2007).

Although indirect, relational and social aggressions have many common elements, certain nuances differentiate one from the other. In indirect aggression the aggressor remains hidden and tries to harm the other either in an undercover manner by, for example, gossiping, spreading rumours or inciting the members of the group to exclude him/her, or physically, by wrecking or stealing his/her property. Relational aggression is characterized by acts that harm the individual's social relations, circle of friends, etc. Finally, social aggression aims to harm the self-esteem and social status of the person attacked (Archer, 2001; Coyne, Archer, & Eslea, 2006). Despite these differences, they tend to be grouped under the term indirect aggression, so this is the term that we shall use here bearing in mind that it refers to the three types described above.

The initial research carried out on this type of aggression considered it to be typically «feminine» and that men showed a greater tendency to commit physical aggression. Numerous studies have demonstrated that men show higher levels of physical aggression than women (see, for example, Archer's meta-analysis, 2004), differences that are present from childhood to elderly (Morales-Vives & Vigil-Colet in press), and this has been shown to

Fecha recepción: 27-4-10 • Fecha aceptación: 7-9-10
Correspondencia: Andreu Vigil-Colet
Centre de Recerca en Avaluació i Mesura de la Conducta
Universidad Rovira i Virgili
43007 Tarragona (Spain)
e-mail: andreu.vigil@gmail.com

be due to true differences and not to measurement instrument bias (Condon, Morales-Vives, Ferrando, & Vigil-Colet, 2006). Although a variety of studies have demonstrated higher levels of indirect aggression in women, others have found no significant differences, especially in adults (Archer, 2004). So although there is sufficient evidence to suggest that in childhood girls have higher levels of indirect aggression than boys, it seems that in adulthood both sexes use this type of aggression equally. All this seems to imply that the differences in indirect aggression that have been attributed to sex seem to reflect the different rates at which boys and girls socialize, and that in adulthood levels of indirect aggression are the same (Lagerspetz & Björkqvist, 1994).

One of the main factors that explains the spate of interest in this type of aggression is that it occurs frequently, particularly in comparison to physical aggression. In this regard, it seems that this type of aggressive behaviour receives less social reprobation than the direct type which prompts adults to channel their aggressiveness by this means. In fact, authors such as Björkqvist (1994) are of the opinion that this type of aggression predominates in adulthood. Furthermore, indirect aggression seems to play a key role in processes of great social repercussion like bullying or mobbing (Björkqvist, Österman, & Hjelt-Bäck, 1994; Björkqvist, Österman, & Lagerspetz, 1994; Coyne, Archer, & Eslea, 2006; Garandeau & Cillessen, 2006).

Although a great deal of research points to the importance of this type of aggression, problems of assessment and measurement limit the number of instruments available for this purpose. In particular, these problems are due to the fact that the subtlety of such behaviours makes them far more difficult to assess than the direct type of aggression. Also, as Forrest, Eatough, & Shevlin (2005) point out, although there are some instruments for assessing indirect aggression in children and adolescents very few have been designed for adults (with the exception of the scales specific for the work place), which makes it more difficult to comprehend and assess this type of aggression in adults. And while a series of consistent predictors have been established for direct aggression —such as impulsivity or deficits in social problem solving— hardly any studies have been made about the predictor variables of indirect aggression.

Many of the instruments for assessing indirect aggression are not specific tests; rather they are subscales of general aggression tests that do not analyse its component elements. This is the case, for example of one of the first questionnaires developed in this field: the direct/indirect aggression scale by Björkqvist, Österman, & Lagerspetz (1992), one of the few questionnaires that has been adapted to Spanish (Toldos, 2005). Other authors have taken this scale as a starting point and have tried to develop scales that assess indirect aggression, relational aggression and social aggression. This is the case of the indirect/social/relational aggression scale by Coyne et al., (2006). Nevertheless, most of these scales have been developed to analyse this type of aggression framework in children and adolescents, and the structure of indirect aggression in adults has only been studied in the workplace (Richardson & Green, 1999).

In this context, we felt that the Indirect Aggression Scales (IAS) specifically developed for adults by Forrest et al., (2005) were particularly promising. These scales introduced two new aspects that should be emphasised. First, the scales had two versions (aggressor and target), which provide a measurement of an individual's tendency to practise this type of aggression or suffer it. Second, they were developed only with items of indirect aggression, unlike other scales that mixed items of both direct and indirect aggression.

When Forrest et al., analysed the factorial structure, they found a three-factor structure for both versions, comprising items of social exclusion, guilt induction and malicious humour. Nevertheless, we consider that there are some methodological limitations that may question the dimensionality of these scales. Firstly, to determine the number of retained factors they used the Kaiser rule (1970), which tends to overestimate the number of factors. Furthermore, the extraction was carried out using a Pearson correlation matrix when polychoric correlation matrixes are more advisable when factorizing items in a Likert response format. Secondly, they applied an orthogonal rotation procedure although it is difficult to assume that the different forms of indirect aggression are independent. In this regard, their loadings matrix reveals that many items showed high loadings on two or more factors.

Taking these limitations into account, the present study aims, first, to make a Spanish adaptation of the indirect aggression scales for target (IAS-t) and aggressor (IAS-a) and determine their dimensionality and factorial structure. Secondly, we aim to analyze the relations between indirect aggression, direct aggression, and impulsivity, because of the well-established relationship between impulsivity and aggressive behaviour and within both forms of aggression (Card, Stucky, Sawalani, & Little, 2008; Vigil-Colet, Morales-Vives, & Tous, 2008). The analysis of the relationships between indirect aggression scales, and aggression scales and impulsivity will be used as an indicator of the convergent and divergent validity of IAS because it is assumed that direct aggression and impulsivity will be related with the aggressor form of IAS but not with the target form.

Finally, we aim to use IAS scores to verify the hypothesis stated above that in adulthood there are no differences in indirect aggression due to sex.

Method

Participants

The participants were 935 university students (434 men and 501 women) aged between 17 and 50 years old (mean= 23.47; standard deviation= 6.74), belonging to different faculties of the Rovira and Virgili University, Tarragona (Spain)

Instruments

Indirect Aggression Scales: The scales proposed by Forrester et al., 2005 were adapted to Spanish using the back-translation procedure described by Hambleton (2005). Two members of the Language Service of the Rovira i Virgili University, with previous experience in adapting psychological tests, made the translations. First, a native Spanish speaker translated the original tests from English to Spanish, and then a native English speaker translated this text back into English. Finally, the back translated version and the original version were compared, and no lack of equivalence was found. Table 1 shows the resulting items of the Spanish version of IAS-a and IAS-t.

Dickman's Impulsivity Inventory (IID). We used the Spanish adaptation of this inventory (Chico, Tous, Lorenzo-Seva, &Vigil-Colet, 2003). It consists of two scales: functional impulsivity (IF) and dysfunctional impulsivity (DF) with reliabilities of 0.78 and 0.76, respectively. Its factorial structure is equivalent to the original English version.

Table 1
ITEMS of the IAS-a and IAS-t scales, item loadings, descriptive statistics, and item-total scale correlations (r_{it})

Test	Item	Loading	Mean	s.d.	r_{it}
IAS-a	He utilizado mi relación con otros para intentar que cambien una decisión	.45	2.14	0.95	.34
	He utilizado el sarcasmo para insultarlos	.50	2.19	1.03	.50
	He intentado influenciarles para que se sintieran culpables	.60	1.74	0.89	.50
	Les he ocultado información que el resto del grupo sabía	.47	1.85	0.87	.37
	Les he excluido de actividades adrede	**.71**	1.37	0.64	.52
	He hecho que los demás no les hablaran	**.81**	1.08	0.33	.41
	Les he excluido de un grupo	**.73**	1.13	0.39	.42
	Me he aprovechado de sus sentimientos para coaccionarles	**.68**	1.26	0.58	.42
	He hecho comentarios despectivos sobre su aspecto	.54	1.74	0.84	.48
	He utilizado bromas privadas para excluirles	**.71**	1.31	0.63	.49
	Les he hecho chantaje emocional	.58	1.50	0.76	.45
	Les he imitado delante de otras personas	.54	1.81	0.92	.45
	He hecho correr rumores sobre ellos	.62	1.22	0.55	.33
	Les he hecho una broma pesada	.57	1.44	0.72	.36
	He hecho algo para que parecieran estúpidos	**.75**	1.27	0.56	.57
	He simulado estar dolido / enfadado con ellos para que se sintieran mal	.55	1.56	0.73	.39
	Les he hecho sentir que no encajaban	**.77**	1.24	0.53	.52
	He hecho que pasaran vergüenza delante de otros	.63	1.27	0.56	.47
	He dejado de hablarles	.49	1.67	0.86	.35
	Les he sometido a presiones innecesarias	.66	1.28	0.58	.39
	Les he excluido de conversaciones adrede	**.71**	1.27	0.57	.54
	Me he burlado de ellos en público	**.67**	1.31	0.62	.51
	Les he insultado	.57	1.56	0.78	.45
	Les he criticado en público	.59	1.71	0.83	.55
	He puesto otras personas en su contra	.77	1.22	0.55	.53
IAS-t	Han hecho que los demás no me hablen	.55	1.43	0.72	.39
	Me han ocultado información que el resto del grupo sabía	.57	1.84	0.86	.46
	Me han hecho pasar vergüenza delante de otros	.60	1.65	0.79	.52
	Me han excluido de un grupo	.61	1.30	0.63	.44
	Me han insultado	.66	1.49	0.77	.52
	Han dejado de hablarme	**.68**	1.42	0.68	.51
	Han utilizado su relación conmigo para intentar que cambie una decisión	.58	1.63	0.84	.48
	Se han aprovechado de mis sentimientos para coaccionarme	**.73**	1.47	0.76	.55
	Se han burlado de mí en público	**.72**	1.31	0.61	.54
	Han simulado estar dolidos y/o enfadados conmigo para que me sintiera mal	**.71**	1.50	0.77	.59
	Han puesto a otras personas en mi contra.	**.74**	1.49	0.75	.63
	Me han hecho sentir que no encajaba	.58	1.51	0.75	.48
	Han hecho correr rumores sobre mí	.65	1.55	0.77	.49
	Me han hecho chantaje emocional	.66	1.52	0.76	.53
	Me han criticado en público	**.73**	1.41	0.67	.56
	Han utilizado bromas privadas para excluirme	**.70**	1.24	0.57	.47
	Me han sometido a presiones innecesarias	.58	1.60	0.82	.32
	Han utilizado el sarcasmo para insultarme	.66	1.34	0.65	.43
	Me han hecho una broma pesada	.56	1.56	0.82	.34
	Han hecho comentarios despectivos sobre mi aspecto	.63	1.34	0.67	.39
	Me han excluido de conversaciones adrede	**.69**	1.27	0.55	.54
	Me han imitado delante de otras personas	.57	1.37	0.66	.41
	Me han excluido de actividades adrede	**.71**	1.25	0.53	.54
	Han hecho algo para que pareciera estúpido	**.67**	1.25	0.56	.48
	Me han intentado influenciar para que me sintiera culpable	.72	1.46	0.70	.59

Buss and Perry Aggressiveness Questionnaire: We used the reduced Spanish version of the questionnaire (Vigil-Colet, Lorenzo-Seva, Codorniu-Raga, & Morales, 2005), consisting of four scales; physical aggression (PA), verbal aggression (VA) anger (AN) and hostility (HO) with reliabilities of 0.92, 0.75, 0.79 and 0.75, respectively. This adaptation presents a good fit to the four-factor model proposed initially by Buss and Perry (1992) and is free of sex bias (Morales-Vives, Codorniu-Raga, & Vigil-Colet, 2005; Condon et al., 2006).

We analysed the data using SPSS 17.0 and FACTOR (Lorenzo-Seva & Ferrando, 2006). We used FACTOR in addition to SPSS for Exploratory Factor Analysis (EFA) because it enabled us to use polychoric correlation matrices and make complementary analyses such as parallel analysis (Horn, 1965).

Procedure

Two professional psychologists administered the tests to groups of between 15-30 individuals in their classrooms. Each individual was randomly assigned one of the scales to answer: IAS-a or IAS-t. In addition, 220 individuals answered the AQ questionnaire and DII. There were two main reasons for applying only one of the IAS forms to individuals. The first was that, in an applied setting, psychologists will probably be interested in one of the two forms of IAS to assess a possible aggressor or a victim of indirect aggression, so it is advisable to analyze the psychometric properties of IAS in the same situation, when only one form is administered. The second reason was that IAS-a and IAS-t are made up of almost the same items, varying only if the individual is the aggressor or the target. In this situation the administration of both forms may introduce carry-over effects, which may disturb subsequent statistical analysis.

Data analysis

Data analysis was performed in two steps. In the first one we analyzed the factorial structure of IAS-a and IAS-t. Taking into account the lack of multivariate normality that is usually related to Likert-type items we used specific methods (Unweighted Least Squares as the extraction method and the polychoric correlation matrix). The dimensionality of the inventories was assessed using parallel analysis. In the second step we analyzed the psychometric properties (reliability and convergent validity) of IAS and sex effects on IAS scores.

Results

Before carrying out the EFA, we computed the values of the Kaiser-Meyer-Olkin index, which were .91 and .93 for IAS-a and IAS-t, respectively, indicating that the correlation matrixes were suitable for factor analysis. The multivariate kurtosis coefficients were 994 and 949, and the corresponding significance tests (Z= 79.1 and 74.6 $p<0.01$) indicated that the multivariate distribution significantly deviated from a normal multivariate distribution. In this situation, a factor analysis method that assumes normal multivariate distribution is not advisable. For this reason we used Unweighted Least Squares (ULS) as the factor extraction method. Furthermore, in this case the Pearson correlation matrix was not appropriate either so we performed EFA on the polychoric correlation matrix (Muthén & Kaplan, 1985; 1992).

The scree tests (Cattell, 1966) shown in Figure 1 suggested that both scales were one-dimensional. The variance accounted for by these factors was 42.34% and 45.21% for IAS-a and IAS-t, respectively. Parallel analysis (Lattin, Carroll, & Green, 2003) was also computed and the dimensionality for both scales proved to be

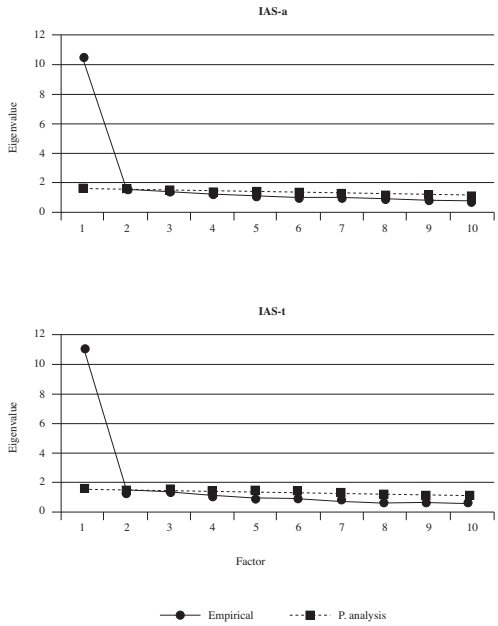


Figure 1. *Scree-test and parallel analysis for aggressor and target versions of IAS*

the same because the eigenvalue of the second factors were below the values that would be expected by chance.

Table 1 shows descriptive statistics for the items of IAS-a and IAS-t and their factorial loadings. As can be seen, all the loadings are greater than 0.40 and the item-total correlations fall in the .30 - .60 interval so there is no need to remove any items because of low loadings or inappropriate item-total relationship. Nevertheless, taking into account their good psychometric properties, 25 items may be excessive for a one-dimensional questionnaire so we also developed a short version of the scales by selecting the 10 items with highest loadings on IAS-a and IAS-t. Table 2 shows descriptive statistics for the full IAS-a and IAS- t scales and for the short scales for men and women. A group of t-tests showed that none of the sex differences was significant either for the full ($t_{(425)}$= 1,333 p>0.05 and $t_{(450)}$= 1,58 p>0.05 for aggressor and target forms respectively) or the short scales ($t_{(425)}$= 1,46 p>0.05 and $t_{(450)}$= 0,13 p>0.05).

Table 3 shows the reliabilities (Cronbach's alpha) for the full and short scales. As can be seen, all the scales showed high reliabilities and the use of short forms did not lead to any significant decrease, so short forms may be a good alternative to the scales initially proposed.

Table 4 shows the Pearson product-moment correlations between the IAS scales, AQ and DII. The aggressor version of IAS showed significant and moderate relationships with all aggression scales and with dysfunctional impulsivity. As was expected, IAS-t had no relationship with impulsivity and aggression measures with the exception of a slight but significant relationship with the AQ hostility scale.

Table 2
Descriptive statistics of full and short IAS versions for men, women and overall sample

	IAS - a		IAS - as		IAS - t		IAS - ts	
	Mean	**S. D.**	**Mean**	**S. D.**	**Mean**	**S. D.**	**Mean**	**S. D.**
Men	37.05	10	13.02	3.85	31.11	8.10	13.81	4.66
Women	35.37	7.91	12.33	3.10	32.15	8.24	14.05	4.11
Total	35.89	8.63	12.56	3.34	31,73	8,18	13.97	4.45

Table 3
Reliabilities (a) and 95% confidence interval for full and short versions of IAS-a and IAS-t

IAS-a	.875	.855 – .893
IAS-as	.818	.788 – .845
IAS-t	.898	.885 – .910
IAS-ts	.849	.830 - .867

Table 4
Product moment correlations between indirect aggression, impulsivity and Aggression Questionnaire scales

	Functio-nal	**Dysfunc-tional**	**Physical**	**Verbal**	**Anger**	**Hostility**	**AQ Total**
IAS-a	.102	**.301**	**.342**	**.325**	**.286**	**.359**	**.462**
IAS-t	-.080	.048	.122	.048	-.012	*.196*	.179

p<0.01; *p<0.05*

Discussion

The results reported above show that the Indirect Aggression Scales provide a reliable measure of indirect aggression in Spanish from both the aggressor and target perspectives. The most prominent difference between our results and the initial proposal made by Forrest et al., (2005) is the dimensionality of the scales. Scree-test and parallel analysis showed that our data has a quite clear one-dimensional structure for both IAS-a and IAS-t. Forrest et al., however, proposed a three-factor structure comprising social exclusion, guilt induction and malicious humour. As we have pointed out above, the methodology they used may have led them to extract too many factors (for example, Kaiser's rule often overestimates the number of retained factors and polychoric correlations are more advisable than Pearson's correlation matrix). The orthogonal loadings matrix of their factorial solution shows that many items have complex loadings, which may indicate that they share a common factor. However, the authors do not provide the correlation matrix between the resulting scales so it is not possible to test this hypothesis. Furthermore, the variance accounted for by their three factors was above 45 per cent, which is the same amount of variance accounted for by our one-dimensional solution.

A possible explanation of the differences between both studies may be that one of the three factors proposed by Forrest et al., accounted for much more variance than the remaining two factors thus giving the scree test the shape of a one-dimensional solution. Nevertheless, this cannot be the case in our study, because the loadings of all items on the factor of indirect aggression are quite similar across the three kinds of items that are supposed to reflect social exclusion, guilt induction and malicious humour. In fact, the ten highest loadings are on a mixture of items.

Another source of evidence that suggests that the scales are one dimensional is that in other indirect aggression scales, such as the Direct and Indirect Aggression Scales (Björkqvist et al., 1992; Toldos, 2005), the Indirect / Social / Relational Aggression Scale (Coyne et al., 2006), or the EXPAGG scale (Tapper & Boulton, 2000) the structure of indirect aggression items was also one-dimensional. Taking all this into account, it seems that indirect aggression items reflect the variability of one latent variable related to indirect aggression and not three independent (orthogonal) latent variables. Nevertheless, further studies in new samples are needed to verify this.

As other studies have shown (for a revision see Archer, 2004), it seems that there are no sex differences in indirect aggression in adulthood, at least at the age range of this study. Nevertheless, future studies with elderly and non university samples will have to generalise this lack of difference because of the specificity of the sample used here. On the other hand, and alternative explanation to this lack of differences is that sex bias in the IAS may be hiding true sex differences so, future research would have to assess the absence or presence of this effect.

The relationships between IAS scales, AQ and DII give the first evidence of IAS validity. Various studies have shown that direct and indirect aggression are related: that is, aggressive individuals seem to present both kinds of aggression (Toldos, 2005; Card et al., 2008). Therefore, a measure of indirect aggression such as IAS-a should be related to a measure of direct aggression such as AQ, which is the kind of relationship found in our study. What is not clear is which determinants make aggressive individuals use direct or indirect aggression. In this regard, situational factors may be

key to understanding why individuals use one form of aggression or another and further research using such methods as three-way component analysis (which take into account situational aspects in psychometric measures) may be helpful (Lorenzo-Seva, Morales-Vives, & Vigil-Colet, 2010).

On the other hand, IAS-t does not show the same pattern of relationships with direct aggression measures, which is what was expected taking into account that it is a measure of suffering aggression not a measure of aggressive behaviour. The only relationship found was with the Hostility scale of AQ which may be explained by the fact that this scale measures a mixture of resentment and mistrust and it seems logical for people who have been suffering aggression to have increased levels of resentment.

IAS-a showed a significant relationship with dysfunctional impulsivity. This is important information because, as Vaillancourt (2005) pointed out, very few studies have examined indirect aggression correlates and there is no previous evidence of relationships between impulsivity and indirect aggression. Many studies have shown that impulsivity, and more specifically dimensions such as dysfunctional impulsivity, highly associated with inhibition deficits are related to direct aggression (Barrat, 1991, 1994; Vigil-Colet et al., 2008). The existence of a positive relationship between impulsivity and indirect aggression seems to show that impulsivity is not only related to primary forms of aggression such as impulsive aggression but also to more sophisticated and less immediate forms of aggression such as indirect aggression. From this viewpoint, impulsivity seems to be a predictor of all forms of aggression and not just specific forms of aggression.

Acknowledgments

This research was supported by a grant from the Spanish Ministry of Education and Science (PSI2008-00236/PSIC).

References

Andreu, J.M., Peña, M.E., & Graña, J.L. (2002). Adaptación psicométrica de la versión española del cuestionario de agresión. *Psicothema, 14,* 476-482.

Archer, J. (2001). A strategic approach to aggression. *Social Development, 10,* 267-271.

Archer, J. (2004). Sex differences in real-word settings: A meta-analytic review. *Review of General Psychology, 8,* 291-332.

Archer, J., Kilpatrick, G., & Bramwell, R. (1995). Comparison of two aggression inventories. *Aggressive Behaviour, 21,* 371-380.

Archer, J., & Coyne, S.M. (2005). An integrated review of indirect, relational and social aggression. *Personality and Social Psychology Review, 9,* 212-230.

Björkqvist, K. (1994). Sex differences in physical, verbal and indirect aggression: A review of recent research. *Sex Roles, 30,* 177-188.

Björkqvist, K., Osterman, K., & Kaukiainen A. (1992). The development of direct and indirect strategies in males and females. In K. Bjorkqvist & P.Niemela (Eds.), *Of mice and women: Aspects of female aggression* (pp. 51-64). San Diego, CA: Academic Press.

Björkqvist, K., Österman, K,. & Hjelt-Bäck, M. (1994). Aggression among University Employees. *Aggressive Behavior, 20,* 173-184.

Björkqvist, K., Österman, K., & Lagerspetz K.M.J. (1994). Sex differences in covert aggression among adults. *Aggressive Behavior, 20,* 27-33.

Björkqvist, K., Lagerspetz K.M.J., & Österman, K. (1992). The direct and indirect aggression scales (DIAS). Vasa, Findland: Abo Academi University, Department of Social Sciences.

Buss, A.H., & Perry, M.P. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology, 63,* 452-459.

Cangas, C., Gázquez, J.J., Pérez-Fuentes, M.C., Padilla, D., & Miras, F. (2007). Evaluación de la violencia escolar y su afectación personal en una muestra de estudiantes europeos. *Psicothema, 19,* 114-119.

Card, N.A., Stucky, B.D., Sawalani, G.M., & Little, T.D. (2008). Direct and indirect aggression during childhood and adolescence: A meta-analytic review of gender differences, intercorrelations and relations to maladjustment. *Child Development, 79,* 1185-1229.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Chico, E., Tous, J.M., Lorenzo-Seva U., & Vigil-Colet A. (2003). Spanish adaptation of Dickman's impulsivity inventory, its relationship to Eysenck's personality questionnaire. *Personality and Individual Differences*, *35,* 1883-1892.

Condon, L., Morales-Vives, F., Ferrando, P.J., & Vigil-Colet, A. (2006). Sex differences in the full and reduced versions of the aggression questionnaire: A question of differential item. *European Journal of Psychological Assessment, 22,* 92-97.

Coyne, S.M., Archer, J., & Eslea, M. (2006). «We're not friends anymore! unlessy»: The frequency and harmfulness of indirect, relational and social aggression. *Aggressive Behavior, 32,* 294-307.

Forrest, S., Eatough, V., & Shevlin, M. (2005). Measuring adult indirect aggression: The development and psychometric assessment of the indirect aggression scales. *Aggressive Behavior, 31,* 84-97.

Garandeau, C.F., &Cillessen, A.N.H. (2006). From indirect agresion to invisible aggression: A conceptual view on bullying and peer group manipulation. *Aggression and Violent Behavior, 11,* 612-625.

Hambleton, R.K. (2005). Issues, designs and technical guidelines for adapting tests into multiple languages and cultures. In R.K.Hambleton, P.F. Merenda y C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). London. L.E.A.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185.

Kaiser, H.F. (1970). A second-generation little jiffy. *Psychometrika 35*, 401-415.

Lagerspetz, K.J.M., & Björkqvist, K. (1994). Indirect aggression in boys and girls. In L.R. Huesmann (Ed.), *Aggressive behaviour: Current perspectives* (pp. 131-150). New York. Plenum.

Lattin, J., Carroll, D.J., & Green, P.E. (2003). *Analyzing multivariate data* (pp. 114-116). Pacific Grove. Duxbury Press.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments and Computers*, *38*(1), 88-91.

Lorenzo-Seva, U., Morales-Vives, F., & Vigil-Colet, A. (2010). Aggressive responses to troubled situations in sample of adolescents: A three-way model approach. *Spanish Journal of Psychology, 13,* 178-189.

Morales-Vives, F., Codorniu-Raga, M.J., & Vigil-Colet, A. (2005). Características psicométricas de las versiones reducidas del cuestionario de agresividad de Buss y Perry. *Psicothema, 17,* 96-100.

Morales-Vives, F., & Vigil-Colet, A. (2010). Are there sex differences in physical aggression in the elderly? *Personality and Individual Differences, 49,* 659-662.

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.

Muthén, B., & Kaplan D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45,* 19-30.

Richardson, D.R., & Green, L.R. (1999). Social sanction and threat explanations of gender effects on direct and indirect aggression. *Aggressive Behavior, 25*, 425-434.

Rodríguez, J.M., Peña, E., & Graña, J.L. (2002). Adaptación psicométrica de la versión española del Cuestionario de Agresión. *Psicothema, 14,* 476-482.

Santisteban, C., Alvarado, J.M., & Recio, P. (2007). Evaluation of a Spanish version of the Buss ansd Perry agression questionnaire: Some personal and situational factors related to the aggression scores of young subjects. *Personality and Individual Differences, 42,* 1453-1462.

Tapper, K., & Boulton, M.J. (2000). Social representations of physical, verbal and indirect aggression: Age and sex differences. *Aggressive Behavior, 26,* 442-545.

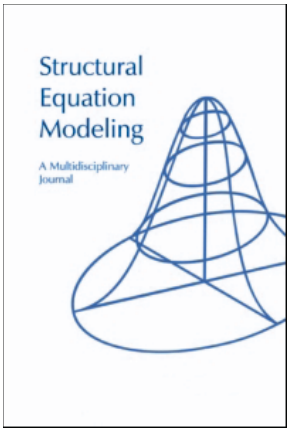
Toldos, M.P. (2005). Sex and age differences in self-estimated physical, verbal and indirect aggression in Spanish adolescents. *Aggressive Behavior, 31,* 13-23.

Tremblay, R.E., & Nagin, D.S. (2005). The developmental origins of physical aggression in humans. In R.E. Tremblay, W.W. Hartup & J. Archer (Eds.), *Developmental origins of aggression* (pp. 83-106). New York. Guilford Press.

Vaillancourt, T. (2005). Indirect aggression among humans: Social construct or evolutionary adaptation? In R.E. Tremblay, W.W. Hartup, & J. Archer (Eds.), *Developmental origins of aggression* (pp. 158-177). New York. Guilford Press.

Vigil-Colet, A., Lorenzo-Seva, U., Codorniu-Raga, M.J., & Morales, F. (2005). Factor structure of the aggression questionnaire among different samples and languages. *Aggressive Behavior, 31,* 601-608.

Vigil-Colet, A., Morales-Vives, F., & Tous, J. (2008). The relationships between functional and dysfunctional impulsivity and aggression across different samples. *Spanish Journal of Psychology, 11,* 480-487.



Structural Equation Modeling: A Multidisciplinary Journal
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653699

A Structural Equation Model at the Individual and Group Level for Assessing Faking-Related Change

Pere Joan Ferrando[a]; Cristina Anguiano-Carrasco[a]
[a] Research Centre for Behavioural Assessment Rovira i Virgili University, Spain

Online publication date: 07 January 2011

To cite this Article Ferrando, Pere Joan and Anguiano-Carrasco, Cristina(2011) 'A Structural Equation Model at the Individual and Group Level for Assessing Faking-Related Change', Structural Equation Modeling: A Multidisciplinary Journal, 18: 1, 91 — 109
To link to this Article: DOI: 10.1080/10705511.2011.532725
URL: http://dx.doi.org/10.1080/10705511.2011.532725

Structural Equation Modeling, 18:91–109, 2011

ISSN: 1070-5511 print/1532-8007 online
DOI: 10.1080/10705511.2011.532725

Psychology Press
Taylor & Francis Group

A Structural Equation Model at the Individual and Group Level for Assessing Faking-Related Change

Pere Joan Ferrando and Cristina Anguiano-Carrasco
Research Centre for Behavioural Assessment Rovira i Virgili University, Spain

This article proposes a comprehensive approach based on structural equation modeling for assessing the amount of trait-level change derived from faking-motivating situations. The model is intended for a mixed 2-wave 2-group design, and assesses change at both the group and the individual level. Theoretically the model adopts an integrative approach that relates the 2 main current conceptualizations of faking, and models the amount of trait change as an individual-differences variable. The model and procedures are used in an empirical study based on 512 participants. Some of the results are interesting and warrant further research. Overall, the methodology that is proposed provides new resources for the theoretical and applied assessment of faking. In particular, it provides the practitioner with new tools for clearly assessing faking at the individual level.

A series of meta-analyses reported in the 1990s (e.g., Barrick & Mount, 1991; Salgado, 1997) revealed that some personality variables could be useful as predictors or moderators of job performance. As a result, in the last two decades interest in the use of personality measures in selection has revived (Morgeson et al., 2007). This interest, in turn, has reopened old debates about the weaknesses of personality measures, and prompted a great deal of new research on these issues (e.g., Zickar & Gibby, 2006). Of these weaknesses, this article deals with faking good, which can be conceptualized as an individual's deliberate attempt to manipulate or distort responses to psychological instruments to create a positive impression (e.g., Furnham, 1986; McFarland & Ryan, 2000; Zickar & Robie, 1999).

As expected from the preceding discussion, most of the new research on faking has been carried out with practical purposes in mind. The most researched topics so far have been the extent to which individuals can fake, the extent to which faking is relevant in high-stakes assessment, and the effects of faking on validity (e.g., Dilchert, Ones, Viswesvaran, & Deller, 2006; Holden, 2006). More substantively oriented research has had less importance, and has mainly been concerned with the conceptualization of faking, the modeling of its mechanisms,

Correspondence should be addressed to Pere Joan Ferrando, Universidad Rovira i Virgili, Facultad de Psicología, Carretera Valls s/n, 43007 Tarragona, Spain. E-mail: perejoan.ferrando@urv.cat

and the role of individual differences (Lautenschlager, 1986; McFarland & Ryan, 2000, 2006; Mersman & Shultz, 1998; Mueller-Hanson, Heggestad, & Thornton, 2006; Zickar & Robie, 1999). This study is more oriented toward this second type of research.

The literature relevant to this article focuses on two general issues. The first is the conceptualization of faking and its operational definition. The second is the designs and procedures that have been used for assessing faking. As for the first point, faking has been operationalized in two main ways (e.g., Griffith & Peterson, 2008): (a) scores on a social desirability scale, and (b) change or difference scores in a measure administered under neutral and faking-motivating conditions. In the first case, the impression management (IM; Paulhus, 1991) component of social desirability has attracted the most interest (McFarland & Ryan, 2006; Mersman & Shultz, 1998). IM is a conscious tendency by respondents to tailor their answers so as to create a more positive social image, and the operationalization is based on the further assumption that faking is the result of this propensity to exaggerate the positive characteristics that IM scales measure.

Difference or change scores are probably the standard measure of faking in present studies (Griffith & Peterson, 2008; McFarland & Ryan, 2006). Whereas the assessment of faking via IM scores is, at best, indirect, change scores are a direct measure of distortion that has obvious face validity. However, they suffer from the well-known shortcomings of all the measures of change based on raw pretest-posttest difference scores: They are inherently unreliable, and prone to end (floor and ceiling) effects (e.g., Webster & Bereiter, 1963). We discuss this second point in detail later.

Zickar and Drasgow (1996) and Zickar and Robie (1999) proposed a model that they called "the theta-shift (θ-shift) model," that (at least theoretically) overcomes the limitations of the raw change scores. Their model assumes that, under faking-motivating conditions, respondents are able to respond to the items as if they have a trait value (θ) different from their true value. So, in these conditions, the true θ level is temporarily changed to improve item scores. Note that change is now defined not at the level of the raw scores, but at the level of the latent trait that these scores measure. Unlike raw scores, the "true" θ-shift scores are unbounded (and so free from end effects) and free from measurement error.

Empirical studies that assess the relations between IM scores and change scores in faking-motivated samples do not provide consistent results, but clearly suggest that they cannot be considered equivalent. Griffith and Peterson (2006, 2008) obtained nonsignificant correlations; McFarland and Ryan (2006) obtained weak positive correlations. Finally, Quist, Arora, and Griffith (2007), Mersman and Shultz (1998), and Wrensen and Biderman (2005) obtained weak negative correlations. The lack of evidence linking both approaches (among other things) prompted some authors to consider IM measures as inappropriate indicators of faking (Burns & Christiansen, 2006; Griffith & Peterson, 2008).

This article is based on the θ-shift model discussed earlier, and proposes an integrative approach that goes in the direction outlined by Holden and Book (2009) and McFarland and Ryan (2006), which integrates the two conceptualizations of faking: IM-based and change-based. To start with, we note that the item contents in IM scales mainly reflect approved behaviors that are believed to be unlikely to occur or attitudes and practices that are socially undesirable but common (minor dishonesties, bad thoughts, weaknesses of character, etc.). So, given their item contents, the IM scales are very susceptible to modification under faking-motivating situations, possibly more than those of any scale of personality content (Burns & Christiansen 2006; Griffith & Peterson, 2008; Viswesvaran & Ones, 1999). Furthermore, it

has been explicitly stated that the θ-shift hypothesis is also applicable to IM measures (Zickar & Robie, 1999). For these reasons, we opted to study faking-related changes within the IM scale itself (as discussed in Holden & Book, 2009). In our approach, then, the IM scores obtained under neutral conditions are used to estimate the individual propensity to fake good, and the changes between the responses to the IM scale administered under neutral and faking-motivating conditions are used to estimate the amount of faking-induced θ-shift. This initial approach can be extended in the future to assess changes in personality content scales.

The IM-based conceptualization of faking has mainly adopted an individual-differences (IIDD) approach in which IM scores are assumed to measure a (possibly complex) IIDD variable that has a certain degree of stability and consistency across different measures (e.g., Furnham, 1986). In contrast, the change-based approach has been mainly situational. However, certain authors (Lautenschlager, 1986; McFarland & Ryan, 2000, 2006; Mersman & Shultz, 1998; Mueller-Hanson et al., 2006) explicitly assumed that raw change scores are also indicators of an IIDD variable, and provided initial evidence based on the variance of the change scores.

In this article we adopt a general IIDD approach and assume that both the initial IM levels and the θ-changes are IIDD variables. However, we do not consider both variables as equivalent. Rather, they are related but different. This conceptualization means that (a) we expect to observe different amounts of individual change under the same faking-motivating conditions, and (b) the amount of individual change cannot be totally predicted from the initial propensity levels. Given the discrepancies in the empirical results discussed earlier, we do not attempt in this first study to derive hypotheses concerning the strength and direction of the relation between both variables. Rather, we attempt to obtain an accurate, model-based assessment of this relation. Burns and Christiansen (2006) considered this assessment as a crucial issue in future research about faking.

We turn now to the second main point: the review of the designs and procedures used to assess faking thus far. However, we must first concern ourselves with the units of analysis. Studies have been made at both the scale score level and the individual item score level. The former tend to be older and based on classical test theory. The latter tend to be more recent and based on item response theory (IRT; see, e.g., Ferrando & Anguiano-Carrasco, 2009; Furnham, 1986). We make our proposal at the item level.

The operationalization of faking via change scores leads naturally to a within-subject design, because this design allows change-based scores to be estimated for each individual. However, both within-subjects and between-subject designs have been used to date (e.g., Henry & Raju, 2006). Methodologically, the within-subjects design is superior in that it has greater statistical power and it controls for individual-differences correlates. However, it is susceptible to carryover or retest effects, and also to history and maturation effects (Shadish, Cook, & Campbell, 2002). As Mesmer-Magnus and Viswesvaran (2006) pointed out, these effects can be controlled much more by using a combined within-subjects and between-subject design and adding a control group. This is the approach proposed here. Specifically, we propose a mixed two-wave, two-group (2W2G) situational design (see, e.g., Henry & Raju, 2006). Participants in the experimental group are administered an IM scale in a situation in which their motivation to fake is low (e.g., neutral conditions, standard test instructions) at Time 1 and in a situation in which their motivation to fake good is higher (e.g., applicants in a high-stakes assessment situation, or respondents specifically instructed to fake good) at Time 2. Participants in the

control group are administered the IM scale under low faking-motivating conditions at both Time 1 and Time 2.

Most studies that assess faking are nonmodel based and analyze raw scores (see, e.g., Furnham, 1986, for a review). Model-based analytical approaches are mainly of two types: IRT based and structural equation modeling (SEM) based. A detailed review and discussion of both approaches is provided in Ferrando and Anguiano-Carrasco (2009). As partly discussed earlier, both IRT and SEM approaches are superior to raw analyses in that they control for measurement error and end effects, and they allow the relations between the fallible raw scores and the traits that these scores intend to measure to be assessed. However, it appears that, to date, no IRT or SEM studies have assessed a mixed design such as the one considered here. As for results, finally, the basic ones are quite consistent across the different designs and analytic strategies: When instructed to fake good, or under highly motivating conditions for doing so, respondents are able to substantially modify their scale scores in measures of personality traits, and they do so in a predictable direction (e.g., Griffin, Hesketh, & Grayson, 2004; Henry & Raju, 2006; Hough, 1998; McFarland & Ryan, 2000; Robie, Zickar, & Schmit, 2001; Viswesvaran & Ones, 1999; Zickar & Robie, 1999).

One clear limitation of most of the previous research is that it is largely designed for the group level, so whatever the design and the analysis strategy might be, the main focus is to assess mean differences between groups or between occasions as indicators of group change. However, the conceptualization of faking as an IIDD clearly calls for a reliable assessment of the amount of individual faking-related change. More generally, accurate measures of change at the individual level are greatly needed in personality measurement (Webster & Bereiter, 1963). In applied settings this individual assessment would serve mainly to flag those individuals who tend to modify their responses most under conditions of pressure or motivation for faking.

In this article we propose a model for assessing faking-related change that is based on the integrated theoretical approach discussed earlier. The model is intended for an IM scale that is administered according to the 2W2G design described previously, and it is assessed both at the group and at the individual level. The assessment at group level consists of fitting a conventional structural equation model followed by a relatively simple reparameterization. The structural equation model at this level is fitted in two steps. First the goodness of model-data fit is assessed. Second, provided that the fit is judged to be acceptable, the parameter estimates are interpreted. The assessment at the individual level is also based on a two-step approach that can be considered, using Fiske's (1968) terms, as the "dual" of the group approach. Therefore, the dual of the first step is to assess the fit of each individual response pattern to the model (i.e., person-fit). This serves to detect those inconsistent individuals for whom the model does not hold and, therefore, whose parameter estimates cannot be meaningfully interpreted. The dual of the second step is to interpret the individual parameter estimates: the initial IM level and the amount of change in each individual respondent.

Overall, this article aims to make three types of contributions: theoretical, methodological, and substantive. At the theoretical level we propose an integrative approach. At the methodological level, we propose a group–individual comprehensive model explicitly based on the theoretical approach. As far as we know, the model is new, and includes some measures of individual change that are original. Finally, at the substantive level we present a study that has substantive interest in itself and that is more than a mere illustration of the model we propose.

The study is based on a design that has been recommended but seldom used in faking applied research (Mesmer-Magnus & Viswesvaran, 2006).

MODEL AND RATIONALE

Consider a test made up of n items that aims to measure IM and is administered twice in two equivalent groups. In Group 1 (control), the test is administered in neutral, standard conditions at both Time 1 and Time 2. In Group 2 (experimental), the test is administered in neutral conditions at Time 1 and under faking-motivating conditions at Time 2. The retest interval is the same in both groups.

In both groups the responses to the n items at Time 1 are modeled as indicators of a latent trait or common factor θ_1, and the responses to the same items at Time 2 are modeled as indicators of a common factor θ_2 (see Jöreskog, 1979; Kenny & Campbell, 1989). In each group, therefore, the general model consists of two measurement submodels linked by a structural submodel that models the relations between θ_1 and θ_2.

The measurement submodels will be considered for three types of responses: (a) binary, (b) graded (treated as ordered categorical variables), and (c) graded or more continuous (e.g., graphic scales or parcels) treated as continuous variables. We adopt B. O. Muthén's (1984) comprehensive framework based on underlying response variables, to which the interested reader is referred for further details. In the first two cases, the measures in each measurement submodel are underlying response variables that are assumed to be normally distributed. In the third case they are directly the item scores. In the first two cases, the observed responses are assumed to arise as a result of a step function governed by a single threshold τ_j (binary case), and by $c - 1$ thresholds (graded case, where c is the number of response categories). In particular, for individual i, belonging to group g ($g = 1$ or 2), responding to item j at time k ($k = 1$ or 2) the relation in the binary case (a) is

$$\begin{aligned} x_{ijk}^{(g)} &= 0 \text{ if } x_{ijk}^{*(g)} < \tau_j \\ x_{ijk}^{(g)} &= 1 \text{ if } x_{ijk}^{*(g)} \geq \tau_j . \end{aligned} \tag{1}$$

Now, for the first two cases the measurement submodel is

$$x_{ijk}^{*(g)} = \lambda_j \theta_{ik} + \varepsilon_{ijk}^{(g)}. \tag{2}$$

And for the third case it is

$$x_{ijk}^{(g)} = \mu_j + \lambda_j \theta_{ik} + \varepsilon_{ijk}^{(g)}. \tag{3}$$

Submodels 2 and 3 are strongly invariant. In submodel 2 the loadings λs and the thresholds τs are assumed to be invariant both over time (stationarity) and across groups (invariance). In submodel 3 the invariant measurement parameters are the loadings and the intercept μs (note in Equations 1–3 that these parameters are not indexed by time subscripts or by group superscripts). As Tisak and Meredith (1990) argued, these invariance and stationarity restrictions are almost necessary to make the model practical. Conceptually they arise from the invariance

property of the regression weights in the regression model and the assumption that the item-trait regressions are an attribute of the items (see Jöreskog, 1979; Lord, 1980; Tisak & Meredith, 1990, for further discussions). Finally, we note that, for the reasons discussed by Little (1997) and B. O. Muthén and Lehman (1985) it was not deemed necessary to further impose strict invariance restrictions in the measurement submodels.

Under normality assumptions and with appropriate reparameterization, in the first case, submodel 2 becomes the IRT bidimensional two-parameter normal-ogive model (2PNOM), and in the second case, the bidimensional graded response model (GRM; see, e.g., McDonald, 1999; Takane & de Leeuw, 1987). Here we shall only use the factor analytic parameterization given in Equations 1 to 3. However, from now on, we shall refer to these models as the 2PNOM and GRM, respectively. Submodel 3 is indeed the well-known Spearman model, and, when applied to item or test responses is usually known as the congeneric model (Jöreskog, 1971), the name we use here.

We turn now to the structural submodel, which is what specifically contains the hypotheses of this study. The structural submodel in Group 1 is

$$\theta_{i2}^{(1)} = \gamma\theta_{i1} + \xi_i^{(1)}. \tag{4}$$

Submodel 4 states that the IM trait at Time 2 is a linear additive function of the trait at Time 1 with a random disturbance (ξ) that represents the aggregation of variables that might have affected the trait during the retest interval (Heise, 1969; Kenny & Campbell, 1989). The random structural disturbance (ξ) is assumed to be uncorrelated with θ_1.

The structural submodel in Group 2 is

$$\theta_{i2}^{(2)} = \gamma(\theta_{i1} + \delta_i) + \xi_i^{(2)}. \tag{5}$$

The parameter δ_i is the one that holds most interest. It models the amount of faking-induced change or θ-shift for individual i. So, as discussed earlier, θ_{i1} measures the standing of individual i in the IIDD variable of propensity to fake, and δi measures the amount of faking-induced change. The main advantage of submodel 5 over previous approaches is that faking-related change is directly modeled as a latent variable (see McArdle, 2009), thus avoiding (in principle) the inherent problems of the raw change scores discussed previously.

In the next two sections we discuss how this model can be fitted at the group and at the individual level. Overall, we propose that the model be treated as a random-regressors model (McDonald, 1982) that is fitted in two stages. In the first stage (group level) the measurement and structural parameters are estimated. In the second stage (individual level), provided that the model–data fit is considered acceptable, the measurement and structural estimates are taken as fixed and known values and used to obtain the individual estimates of interest.

STAGE 1: ANALYSES AT THE GROUP LEVEL

The standard structural submodel for a two-wave design is (e.g., Jöreskog, 1979; Kenny & Campbell, 1989; L. K. Muthén & Muthén, 2007):

$$\theta_{i2} = \alpha + \beta\theta_{i1} + \xi_i. \tag{6}$$

The approach we propose at group level is to fit the full structural equation model for the 2W2G with the strongly invariant measurement submodels described earlier and the standard structural submodel in Equation 6. Then we obtain estimates of interest in our alternative formulation of the structural submodel (Equations 4 and 5) by means of a simple reparameterization. We start by fitting the following two-group structural submodels:

$$\begin{aligned} \theta_{i2}^{(1)} &= \beta^{(1)}\theta_{i1} + \xi_i^{(1)} \\ \theta_{i2}^{(2)} &= \alpha + \beta^{(2)}\theta_{i1} + \xi_i^{(2)}. \end{aligned} \tag{7}$$

In both groups we scale θ_1 as a standard variable (zero mean and unit variance), so the means and variances of interest in this section can be identified, estimated, and interpreted relative to these fixed values. Now, a comparison of Equations 4 and 7 shows at once that

$$\hat{\gamma} = \hat{\beta}^{(1)}. \tag{8}$$

Next, by taking expectations from Equations 5, 7, and 8, we obtain the mean θ-shift estimate in the experimental group:

$$\hat{\mu}(\delta) = \frac{\hat{\alpha}}{\hat{\beta}^{(1)}} \tag{9}$$

which measures the overall amount of faking-related change for this group.

The estimated covariance between the initial IM levels and the amount of individual faking in the experimental group is also obtained by using Equations 5, 7, and 8 and taking expectations.

$$\hat{\sigma}(\theta_1, \delta) = \frac{\hat{\beta}^{(2)}}{\hat{\beta}^{(1)}} - 1. \tag{10}$$

Finally, by developing the variances of θ_2 in Equations 4 and 5 and subtracting, we obtain:

$$\frac{\hat{\sigma}^2(\theta_2^{(2)}) - \hat{\sigma}^2(\theta_2^{(1)})}{\hat{\beta}^{(1)^2}} = \hat{\sigma}^2(\delta) + 2\hat{\sigma}(\theta_1, \delta). \tag{11}$$

where the last estimate on the right side of Equation 11 is obtained from Equation 10. Therefore, Equation 11 allows the variance of δ to be estimated. This variance is the model-based counterpart of the raw-change-score-based variance used in previous studies to assess whether faking is an IIDD variable. It therefore measures the relevancy of δ as an IIDD. A low estimate would mean that the role of IIDD is negligible because, under the same conditions, the amount of change is virtually the same for all individuals. As mentioned earlier, the value of the estimate (Equation 11) must be interpreted in relation to the fixed unit variance at Time 1.

Further measures of interest can be obtained from the basic estimates just derived. Thus, to assess the relevancy of the amount of faking-related change at the group level, we can compute Cohen's (1988) *d* effect-size measure as

$$d = \frac{\hat{\mu}(\delta)}{\hat{\sigma}(\delta)}. \tag{12}$$

In the same way, the estimated product-moment correlation between the initial IM levels and the amount of individual faking can now be obtained by

$$\hat{\rho}(\theta_1, \delta) = \frac{\hat{\sigma}(\theta_1, \delta)}{\hat{\sigma}(\delta)}. \tag{13}$$

One final general remark should be made regarding the assessment at this stage. The literature on SEM in which the same variables are measured at two points in time consistently recommends that the errors corresponding to the same indicators should be correlated (e.g., Jöreskog, 1979; Kenny & Campbell, 1989; Pitts, West, & Tein, 1996). In effect, it is reasonable to assume that part of the measurement error in this case is not random, but systematic, reflecting unique item content, memory and other retest effects, or both (Pitts et al., 1996). Furthermore, failure to allow for correlated errors is expected to lead to a decrease in the overall fit of the model as well as to biased parameter estimates (particularly those corresponding to the structural submodel; see Pitts et al., 1996). The strategy we follow in our study is to compare the fit of the models with and without correlated errors, study the changes in the structural estimates of most interest, and decide which is the most appropriate model.

STAGE 2: ASSESSING THE IMPACT OF FAKING AT THE INDIVIDUAL LEVEL

We assume that the structural equation model has been fitted and that model–data fit is acceptable. If this is so, the parameter estimates will be taken as fixed and known values, and used to obtain individual trait estimates in θ_1 and θ_2 (i.e., the estimated factor scores) for the respondents in the experimental group. These factor scores, in turn, will be the basis for the two steps proposed at this stage: assessing individual (person) fit, and obtaining and interpreting the individual parameter estimates of IM and change. In principle, we shall consider the maximum likelihood (ML) point estimates of the individual trait levels in θ_1 and θ_2. For $k = 1$ or 2, we shall denote the ML estimates as $\hat{\theta}_k$, and write them generally as

$$\hat{\theta}_{ik} = \theta_{ik} + \omega_{ik} \tag{14}$$

where θ_{ik} is the true trait level, and ω_{ik} is an error term independent from θ_{ik} (see, e.g., Samejima, 1969). Asymptotically (as the number of items become large) the ML estimates are unbiased (i.e., $E(\hat{\theta}_{ik}|\theta_{ik}) = \theta_{ik}$) and normally distributed, and its variance is given by the inverse of Fisher's information matrix:

$$\mathbf{I}(\theta_{1,}\theta_{2,}) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} = -E \begin{bmatrix} \dfrac{\partial^2 lnL}{\partial\theta_1^2} & \dfrac{\partial^2 lnL}{\partial\theta_1\partial\theta_2} \\ \dfrac{\partial^2 lnL}{\partial\theta_2\partial\theta_1} & \dfrac{\partial^2 lnL}{\partial\theta_2^2} \end{bmatrix} \tag{15}$$

where *ln L* is the logarithm of the likelihood function. Because the model we propose is based on an independent-cluster solution (i.e., each variable has only one nonzero loading on one of

the two factors), the information matrix in our case is diagonal, and the nonnull elements I_{11} and I_{22} are the asymptotic variances of $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$, respectively.

The particular form of the ML point estimates and the corresponding information matrix depend on the particular response model that is fitted. In the 2PNOM and the GRM (see, e.g., Samejima, 1969) the ML equations are nonlinear and the estimates must be obtained iteratively. Furthermore, the information values are generally different at different trait levels. In the congeneric model the ML estimates can be obtained in closed form: They are the well-known Bartlett's weighted least square factor scores (e.g., McDonald, 1982). Furthermore, the information values do not depend on the trait levels (see, e.g., Ferrando, 2009).

As discussed earlier, before the individual estimates are interpreted, the fit of each individual response pattern should be assessed in Step 1. The approach we propose is to assess the fit of each response pattern at Time 1. If a response pattern is inconsistent with the model at Time 1, so the estimated initial trait level cannot be meaningfully interpreted, there is little sense in continuing the analysis to estimate the change produced under faking-motivating conditions. In agreement with the ML estimation, we propose to use an ML-based general class of measures of fit: the likelihood-based person–fit indexes (see, e.g., Meijer & Sijtsma, 1995). The general idea on which these indexes are based is that the likelihood function value of a particular item response pattern given the response model will be large for patterns that are consistent with the model and small for inconsistent patterns. Assuming that the parameters of the model are fixed and known, the basic index is simply the logarithm of the likelihood function evaluated at the maximizing value of θ (i.e., the ML estimate). There are standardized versions of the basic index for the 2PNOM (Levine & Rubin, 1979), the GRM (Drasgow, Levine, & Williams, 1985), and the congeneric model (Ferrando, 2006). These are the versions that we recommend in the approach we propose.

We turn now to Step 2. For those individuals who were considered consistent at Time 1 according to the person–fit analysis, we propose to estimate the amount of θ-shift by using the following statistic:

$$\hat{\delta}_i = \frac{\hat{\theta}_{i2}}{\gamma} - \hat{\theta}_{i1} \tag{16}$$

which is an ML estimate of δ_i in Equation 5. Under the assumptions previously discussed, and given that $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$ are unbiased, it follows that $\hat{\delta}$ is an unbiased estimate of δ or, in other words, an unbiased estimate of the amount of individual θ-shift. The corresponding error variance is estimated by

$$Var(\hat{\delta}_i | \delta_i) = \frac{1}{\gamma^2 I_{22}(\hat{\theta}_{i2})} + \frac{1}{I_{11}(\hat{\theta}_{i1})}. \tag{17}$$

Note that the covariance term is zero because the information matrix is diagonal. The square root of Equation 16 is the standard error of the estimate (*SE*), and can be used for setting confidence intervals around $\hat{\delta}$ as well as for assessing the null hypothesis of zero θ-shift. Hence, using the approximate normality property of the ML estimates, the $(1-\alpha)\times 100$ confidence interval for the amount of individual change is

$$\hat{\delta}_i \pm z_{1-\frac{\alpha}{2}} s.e.(\hat{\delta}_i). \tag{18}$$

The assessment approach in Equations 15 to 17 can be considered as a refined SEM-based version of the idiographic test ratio that Webster and Bereiter (1963) proposed for raw change scores. Now, provided that the individual amount of θ-shift is considered significant, it is of interest to further obtain a scaled measure for assessing the magnitude of this change. The simple measure we propose is

$$zd(\hat{\delta}_i) = \frac{\hat{\delta}_i}{\sqrt{Var(\hat{\theta}_1)}}; \tag{19}$$

where the marginal variance of $\hat{\theta}_1$ can be obtained by developing the variance of the two independent terms in Equation 13. The *zd* measure in Equation 18 is an index of relative position with respect to the group, and assesses the magnitude of individual change in standard deviation units with respect to the marginal distribution of the estimated trait levels at Time 1. Conceptually, it measures how the position of individuals on the Time 1 distribution would have changed if, instead of responding as they did under neutral conditions, they had responded as they did subsequently under faking-motivating conditions.

In the 2PNOM and the GRM, in which the ML estimates cannot be obtained in close form, the estimates corresponding to extreme response patterns are not defined, and those derived from some of the almost-extreme patterns can become very unstable, with unreasonably large values and standard errors (e.g., Zimowski, Muraki, Mislevy, & Bock, 2003). This problem must be taken into account in our case because, under faking-motivating conditions, many patterns are expected to become extreme. If they do, the solution we propose is to use the Bayes modal estimate (MAP), by assuming a prior bivariate normal distribution for θ_1 and θ_2, in which the mean vector and the covariance matrix are those obtained when fitting the model at the group level. Conceptually the MAP estimate can be considered equivalent to the ML estimate discussed so far, but augmented with an additional "item" that contains the prior distributional information just described (Wainer & Mislevy, 2000). Because of this information, the precision of the MAP estimate typically exceeds that of the ML estimate by a positive term that depends on the prior (Wainer & Mislevy, 2000). In our case, the MAP-corrected information elements can be readily obtained as:

$$\begin{aligned} I_{11}(MAP) &= I_{11} + 1 \\ I_{22}(MAP) &= I_{22} + \frac{1}{Var(\theta_2)}. \end{aligned} \tag{20}$$

The standard errors and confidence intervals can be computed according to Equations 16 and 17 using these MAP-corrected information elements.

EMPIRICAL STUDY

Method

Participants and procedure. A total of 512 undergraduates from the psychology and social sciences faculties of a Spanish university took part in the study. The measure described

here was voluntarily administered in classroom groups of 25 to 60 students at two points in time with a retest interval of 6 weeks. In all cases the same person administered the measure in a paper-and-pencil version. The administration was anonymous, and the respondents had to provide only three particulars: gender, age, and favorite color.

At Time 1 all the participants were asked to respond under the standard instructions provided in the manuals of Eysenck's questionnaires. Among other things, these instructions advise giving an honest answer. At Time 2, each classroom group was divided in half, and the half groups were assigned randomly to the two different conditions that define the design groups. The participants assigned to Condition 1 (control group, $n = 235$) were retested using the same standard instructions as at Time 1. Participants assigned to Condition 2 (experimental group, $n = 277$) were given the instructions detailed in S. B. G. Eysenck, Eysenck, and Shaw (1974). Respondents are asked to imagine themselves as job applicants applying for a job that they really wanted. They should try to give a good impression when answering by putting what they think the employer would like them to be regardless of the truthful answer. Overall, the two-group part of the study used a quasi-experimental design (Shadish et al., 2002) in which clusters defined by formal institutions (i.e., classroom groups) were half-split and randomly assigned to the experimental conditions. Although the degree of comparability is never as high as with purely random individual assignment, "quasi-comparability" can safely be assumed in this case. In both groups, respondents are approximately the same age ($M = 19.85$ in Group 1 and 21.54 in Group 2) and have a similar sociocultural background. The proportion of genders is about the same (77% female in Group 1 and 75% in Group 2). Purely individual random assignment was not attempted because of practical and time constraints.

Measures. We used the 21-item Lie scale that is included in the Eysenck Personality Questionnaires for Adults (H. J. Eysenck & Eysenck, 1975; S. B. G. Eysenck, Eysenck, & Barrett, 1985). According to Paulhus (1991), the Lie scales in Eysenck's questionnaires are almost pure measures of IM. Furthermore, the Lie scale has been intensively used in research and the evidence suggests that it is essentially unidimensional and has acceptable psychometric properties (Ferrando, Chico, & Lorenzo, 1997; Furnham, 1986).

The items in the Lie scale are binary, so the measurement submodel we fitted was based on the 2PNOM (first case, Equations 1 and 2). Preliminary analyses showed that, in the experimental group, one of the items had zero variance under the faking-motivating condition. So, to avoid problems when fitting the structural equation model this item was omitted from the analyses from the beginning.

Results

Group-level analyses. The structural equation model with the 2PNOM-based measurement submodels was fitted using weighted least squares with mean correction (WLSM) estimation as implemented in the M*plus* program version 5.1 (L. K. Muthén, & Muthén, 2007). This procedure, based on a simplified weight matrix together with a mean-adjusted chi-square statistic, seems to work well in the case of questionnaires of realistic length and not very large sample sizes, which is the case here. As for the indexes of fit, as well as the adjusted chi-square statistic, we considered the root mean squared error of approximation point estimate (Browne & Cudeck, 1993) and the comparative fit index (CFI; Bentler & Bonett, 1980). As discussed

TABLE 1
Goodness of Fit Results for the Group-Level Structural Equation Model

Model	χ^2	*df*	$\Delta\chi^2$	Δdf	*RMSEA*	*CFI*	ΔCFI
Independent errors	3,235.49	1,555	556.90	40	0.063	0.90	0.03
Correlated errors	2,664.85	1,515			0.050	0.93	

Note. RMSEA = root mean squared error of approximation; CFI = comparative fit index.

previously, the model was fitted with and without retest correlated residuals. The additional measures considered for assessing the differences of fit between the two corresponding nested models were the Satorra–Bentler corrected chi-square difference test (Satorra, 2000) and the ΔCFI using the reference value of .002 proposed by Meade, Johnson, and Braddy (2008). The results are shown in Table 1.

Overall, the results in Table 1 suggest that the fit of the model without correlated residuals is at the very limit of being acceptable, whereas the fit of the model with correlated retest residuals can be considered satisfactory. Furthermore, according to the difference measures discussed earlier, the fit appears to substantially improve when going to the less restricted correlated-residuals model, so this is the model we discuss from this point on. However, it should be pointed out that the structural estimates of most interest in this study were very similar in both models. Finally, we believe that the acceptable fit is a noticeable result, as models of the type considered here are usually intended to make a detailed scrutiny of small sets of 5 to 10 items (B. O. Muthén & Lehman, 1985). However, here we assess a set of 40 binary variables measured simultaneously in two groups.

Table 2 shows the reparameterized structural estimates in the experimental group obtained according to Equations 8 to 13. We first note that the amount of θ-shift at the mean group level is in the expected direction and substantial, and the effect size is large. This result agrees with and reinforces the evidence discussed earlier (in particular Viswesvaran & Ones, 1999) but adds little really new to the existing knowledge. Second, if compared to the unit value used to scale the trait variance under neutral conditions, the estimated variance of δ is very large. So, it appears that the interindividual variability in the amount of change elicited by the faking-motivating conditions is very large. This result justifies the conceptualization of δ as an IIDD variable. Finally, the amount of change seems to be negatively related to the initial trait level. This result is discussed in detail later.

Individual-level analyses. The parameter estimates obtained at the group level were taken as fixed and known and for each respondent (i) we obtained (a) the person–fit lz value at Time 1 (lz_i), (b) the trait estimates at Time 1 and at Time 2 ($\hat{\theta}_{i1}\hat{\theta}_{i2}$), (c) the estimated amount of

TABLE 2
Parameter Estimates of the Structural Submodel

$\hat{\mu}(\delta)$	*Cohen's d*	$\hat{\sigma}^2(\delta)$	$\sigma^2(\theta_1)$	$\hat{\sigma}(\theta_1, \delta)$
2.71	1.39	3.81	1 (fixed)	−0.38

theta change, according to Equation 15 ($\hat{\delta}_i$), and (d) the 90% confidence interval around $\hat{\delta}_i$. For the reasons discussed previously the trait estimates were MAP estimates, so the standard error was obtained using Equation 19.

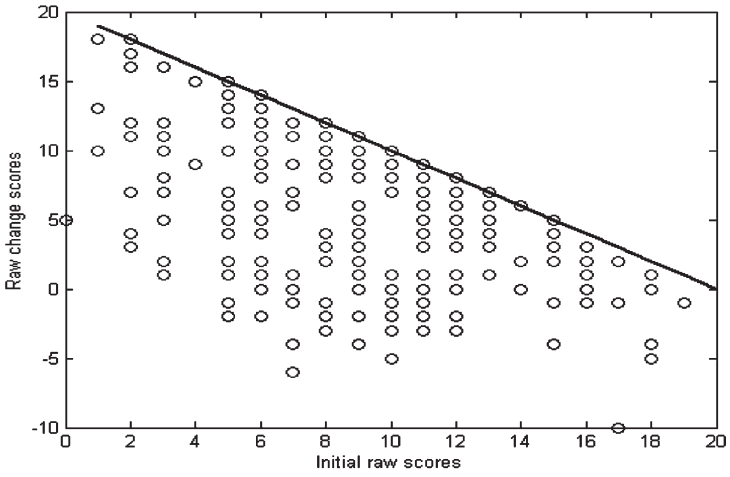
We first illustrate how the preceding estimates can be used for individual assessment by using data from three participants. The point estimates for Respondent 1 were $lz_1 = 0.71$, $\hat{\theta}_{1,1} = 0.06$, and $\hat{\delta}_1 = 4.40$. The 90% confidence interval around $\hat{\delta}_1$ was (2.43, 6.36). These results suggest that Participant 1 responded quite consistently at Time 1 (the *lz* value is positive and fairly large), and that his or her level of IM at Time 1 was about average, which suggests that he or she has a moderate tendency to describe himself or herself in favorable terms under neutral, standard conditions. However, under faking-motivating conditions the θ-shift upward change of this individual was very large. The amount of change is indeed statistically significant, as the zero value is considerably below the lower limit of the confidence interval. The value of the scaled measure of change *zd* in Equation 18 was 3.89. So, if this individual had responded at Time 1 as he or she did under faking instructions, he or she would have been 3.89 *SD* above the initial location.

The point estimates for Respondent 216 were $lz_{216} = 0.47$, $\hat{\theta}_{216,1} = -1.49$, and $\hat{\delta}_{216} = 0.32$. The 90% confidence interval was (−1.14, 1.79). Like Respondent 1, the *lz* value suggests that Respondent 216 responded consistently at Time 1. However, in this case the estimated level at Time 1 is negative and quite low, which suggests that this respondent is not particularly predisposed to describe himself or herself in favorable terms (quite the opposite, in fact). Furthermore, the amount of estimated θ-shift is small, and cannot be considered to be statistically significant because the confidence interval includes the zero value. Overall, the result suggests that the respondent tends to appraise himself or herself critically and even negatively, and does not change this tendency even when he or she has been explicitly instructed to describe himself or herself in favorable terms.

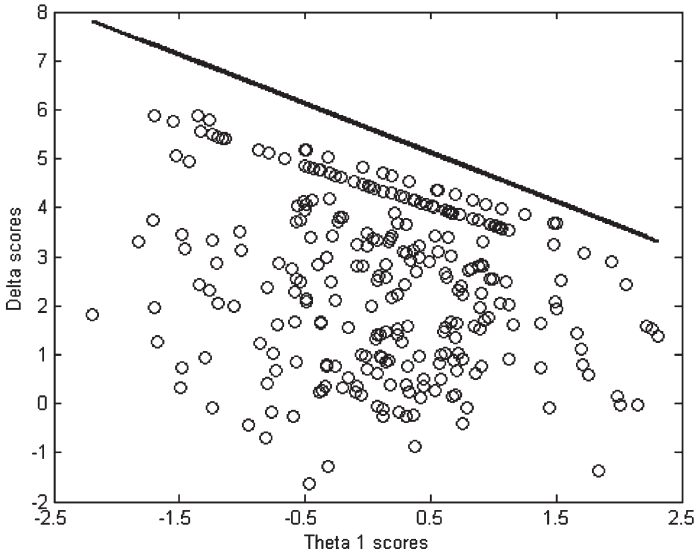
Finally, the point estimates for Respondent 74 were $lz_{74} = -2.26$, $\hat{\theta}_{74,1} = 0.75$, and $\hat{\delta}_{74} = -0.41$. If we attempt to interpret the change for this individual, the negative value of $\hat{\delta}_1$ suggests that under faking-motivating conditions he or she described himself or herself less favorably than under neutral conditions. However, the large negative value of the person–fit statistic suggests that this respondent responded inconsistently at Time 1. So, interpretations about either his or her initial results and further changes are not warranted and should not be attempted.

We now move on to discuss the relations between the initial levels at Time 1 and the amount of change under faking-motivating conditions. The covariance between the individual estimates was −0.32, which is not far from the structural value obtained when the structural equation model was fitted at the group level (see Table 2). The product-moment correlation was $r = -0.22$. Before this result is interpreted, however, some discussion is needed.

If the preceding result had been obtained with raw scores (as in many previous studies), the negative relation between the initial level and the amount of change might well have been an artifact due to ceiling effects. Indeed, for those respondents who at Time 1 already scored near the top of the scale, there is much less room for further increase at Time 2 than for those respondents who at Time 1 attained a lower score. This effect is apparent in Figure 1a, which plots the amount of raw change (obtained as the difference between the raw Lie score at Time 2 minus the raw Lie score at Time 1) against the raw Lie score at Time 1. Figure 1a also shows the line of maximum upward change for each initial score (in this case the line is



(a)



(b)

FIGURE 1 (a) Amount of raw change against the raw Lie score at Time 1. (b) Relation between the trait estimates at Time 1 and the δ estimates.

$Y = 20 - X$). Note how the scatter of points is "compressed" against the line. The product-moment correlation is $r = -0.51$, which is fairly strong. However, at least part of this relation is likely to be due to the ceiling effect. This is one of the limitations of defining faking in terms of change raw scores.

In this study we do not work with raw scores but with trait estimates. As discussed earlier, the "true" trait levels are unbounded, and so free from end (floor and ceiling) effects. In practice, however, the trait estimates are essentially nonlinear transformations of the raw scores that map these scores on the entire real axis (see, e.g., Ferrando, 2002). Thus, the transformation "stretches" the raw scale and so minimizes the end effects, but this does not mean that they are totally removed. Figure 1b plots the relation between the trait estimates at Time 1 and the δ estimates. In this case the line of maximum change cannot be defined because both variables are theoretically unbounded. However, we also plotted a line based on the maximum score estimate at Time 2. In this case, Figure 1b suggests that there is still some room for further upward changes. Indeed, further research is needed on this important issue. However, the results reported here suggest that there seems to be a moderate negative relation between the initial trait levels and the amount of individual faking-related change.

DISCUSSION

This article aimed to make substantive, methodological, and empirical contributions to the assessment of the impact of faking on personality measures and apply these contributions in an empirical study. With respect to the existing literature, it can be considered mainly an extension of the recent contributions made by Ferrando and Anguiano-Carrasco (2009). At the substantive level, the main contribution of this study is the distinction between propensity (initial IM levels) and θ-changes, both conceptualized as IIDDs variables. At the methodological level, our study extends the simple MG design proposed by Ferrando and Anguiano-Carrasco (2009) to the 2W2G design, and generalizes the binary model to the graded and continuous case. The repeated-measures part of the model allows accurate individual parameter estimates of IM and change to be obtained, and this part of the study appears to be new. At the empirical level, finally, as far as we know, this is the first SEM study based on real data that uses a 2W2G design and that is based on the 2PNOM.

Overall, the results of the study suggest that the proposed procedures are feasible and that their use provides meaningful results. To start with, at the mean-group level the results obtained are in agreement with Viswesvaran and Ones's (1999) conclusions, and suggest that the IM measures are easy modifiable and very susceptible to significant change under faking-motivated or instructed conditions. Of greater interest is the large estimated variance of δ relative to the unit fixed trait variance at Time 1. In principle, the large interindividual variability under faking-motivating instructions supports the conceptualization of δ as an IIDD variable, as proposed by Furnham (1986), Lautenschlager (1986), McFarland and Ryan (2000, 2006), Mersman and Shultz (1998), or Mueller-Hanson et al. (2006). The large variability also suggests that this variable might be quite complex. Some of the proponents of change as an IIDD variable (McFarland & Ryan, 2000, 2006; Mersman & Shultz, 1998; Mueller-Hanson et al., 2006) suggested that this variable might jointly reflect a predisposition to deceive and the ability to do so, and that this is the reason for its complexity. They also proposed that antecedent

variables be used so that the two components could be separated. We believe that this is an interesting line of research, and note that the model-based framework we propose is particularly appropriate for it.

A third noteworthy result is that the amount of change seems to be negatively related to the initial trait level, which agrees with previous reports by Quist, Arora, and Griffith (2007), Mersman and Shultz (1998), and Wrensen and Biderman (2005). In previous raw-score-based studies this result might be an artifact due to end effects. In our trait-based study the results suggest that the negative relation is substantive or, at least, cannot be completely explained in terms of end effects. So far no explanation has been provided for this phenomenon. However, the result suggests that those individuals who tend to describe themselves in excessively positive terms even in neutral situations are only slightly affected when they are instructed to exaggerate their positive qualities. In other words, imposing the condition of falsehood has a greater effect on honest individuals.

In our opinion, the methodology we propose is particularly useful for individual assessment. It allows inconsistent respondents to be flagged at Time 1 and, in particular, the amount of individual change to be accurately assessed. This assessment can be useful for practitioners in certain situations, for example, for prison permission evaluators who could compare test scores of prisoners during their imprisonment (when they are not motivated to fake) and just before the permission (very motivated to fake). A second situation could be in long job selection processes, in which the candidates would be much more liable to fake in the final phases after weeks of trials than in the initial phases when hundreds of candidates are interviewed.

The study also has some clear limitations. The research is in its initial stages and can be improved in many areas. In further studies, for example, the design could be strengthened by using a pure individual random assignment (if practically feasible). Also, we used instructed faking, but it would be very interesting to evaluate real selection scenarios or groups trained to deceive. Although we only considered a single measure with a specific item format, we propose a more general procedure that would be better assessed with different measures and item formats. Even more important, the results reported here are only a starting point. The large estimated interindividual variability in change requires further research into the conceptualization, antecedents, and correlates of this (possibly complex) variable. Also, the results concerning the relations between change and the initial IM levels are far from being conclusive and clearly require replication, if possible, with higher ceiling measures.

ACKNOWLEDGMENTS

This research was partially supported by a grant from the Catalan Ministry of Universities, Research and the Information Society (2005SGR00017), and by a grant from the Spanish Ministry of Education and Science (PSI2008–00236/PSIC).

REFERENCES

Barrick, M. R., & Mount, M. K. (1991). The Big-Five personality dimensions and job performance. *Personnel Psychology, 44,* 1–26.

Bentler, P. M., & Bonett, D. G. (1980). Significance test and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

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

Burns, G. N., & Christiansen, N. D. (2006). Sensitive or senseless: On the use of social desirability measures in selection and assessment. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 113–148). Greenwich, CT: Information Age.

Cohen, J. (1988). Set correlation and contingency tables. *Applied Psychological Measurement, 12,* 425–434.

Dilchert, S., Ones, D. S., Viswesvaran, C., & Deller, J. (2006). Response distortion in personality measurement: Born to deceive, yet capable of providing valid self-assessments? *Psychology Science, 48,* 209–225.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67–86.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire.* London: Hodder & Stoughton.

Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. T. (1985). A revised version of the Psychoticism scale. *Personality and Individual Differences, 6,* 21–29.

Eysenck, S. B. G., Eysenck, H. J., & Shaw, L. (1974). The modification of personality and Lie scores by special "honestly" instructions. *British Journal of Social and Clinical Psychology, 13,* 41–50.

Ferrando, P. J. (2002). Theoretical and empirical comparisons between two models for continuous item responses. *Multivariate Behavioral Research, 37,* 521–542.

Ferrando, P. J. (2006). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research, 42,* 481–507.

Ferrando, P. J. (2009). Difficulty, discrimination and information indices in the linear factor-analytic model for continuous responses. *Applied Psychological Measurement, 33,* 9–24.

Ferrando, P. J., & Anguiano-Carrasco, C. (2009). Assessing the impact of faking on binary personality measures: An IRT-based multiple-group factor analytic procedure. *Multivariate Behavioral Research, 44,* 497–524.

Ferrando, P. J., Chico, E., & Lorenzo, U. (1997). Dimensional analysis of the EPQ-R Lie scale with a Spanish sample: Gender differences and relations to N, E, and P. *Personality and Individual Differences, 30,* 641–656.

Fiske, D. W. (1968). Items and persons: Formal duals and psychological differences. *Multivariate Behavioral Research, 3,* 393–401.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7,* 385–400.

Griffin, B., Hesketh, B., & Grayson, D. (2004). Applicants faking good: Evidence of item bias in the NEO PI–R. *Personality and Individual Differences, 36,* 1545–1558.

Griffith, R. L., & Peterson, M. H. (2006). *A closer examination of applicant faking behaviour.* Greenwich, CT: Information Age.

Griffith, R. L., & Peterson, M. H. (2008). The failure of social desirability measures to capture applicant faking behavior. *Industrial and Organizational Psychology, 1,* 308–311.

Heise, D. R. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review, 34,* 93–101.

Henry, M. S., & Raju, N. S. (2006). The effects of traited and situational impression management on a personality test: An empirical analysis. *Psychology Science, 48,* 247–267.

Holden, R. R. (2006, September). *Faking on noncognitive self-report: Seven primary questions.* Paper presented at the ETS Mini-conference on Faking in Noncognitive Assessments, Princeton, NJ.

Holden, R. R., & Book, A. S. (2009). Using hybrid Rasch-latent class modeling to improve the detection of fakers on a personality inventory. *Personality and Individual Differences, 47,* 185–190.

Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11,* 209–244.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal-developmental investigations. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 263–302). New York: Academic.

Kenny, D. A., & Campbell, D. T. (1989). On the measurement of stability in over-time data. *Journal of Personality, 57,* 445–481.

Lautenschlager, G. J. (1986). Within-subject measures for the assessment of individual differences in faking. *Educational and Psychological Measurement, 46,* 309–316.

Levine, M. V., & Rubin, D. F. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269–290.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32,* 53–76.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60,* 577–605.

McDonald, R. P. (1982). Linear vs. nonlinear models in item response theory. *Applied Psychological Measurement, 6,* 379–396.

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

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85,* 812–821.

McFarland, L. A., & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology, 36,* 979–1016.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93,* 568–592.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item scores patterns: A review and new developments. *Applied Measurement in Education, 8,* 261–272.

Mersman, J. L., & Shultz, K. S. (1998). Individual differences in the ability to fake on personality measures. *Personality and Individual Differences, 24,* 217–227.

Mesmer-Magnus, J., & Viswesvaran, C. (2006). Assessing response distortion in personality tests: A review of research designs and analytic strategies. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behaviour* (pp. 85–113). Greenwich, CT: Information Age.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60,* 683–729.

Mueller-Hanson, R. A., Heggestad, E. D., & Thornton, G. C. (2006). Individual differences in impression management: An exploration of the psychological process underlying faking. *Psychology Science, 48,* 288–312.

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

Muthén, B. O., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10,* 133–142.

Muthén, L. K., & Muthén, B. O. (2007). *M*plus *user's guide* (5th ed.). Los Angeles: Muthén & Muthén.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Pitts, S. C., West, S. G., & Tein, J. (1996). Longitudinal measurement models in evaluation research: Examining stability and change. *Evaluation and Program Planning, 19,* 333–350.

Quist, J. S., Arora, S., & Griffith, R. L. (2007, April). *The association of social desirability and applicant response distortion: A validation study.* Paper presented at the 22nd annual Society for Industrial and Organizational Psychology conference, New York.

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14,* 187–207.

Salgado, J. F. (1997). The five factor model of personality and job performance in the European community. *Journal of Applied Psychology, 82,* 30–43.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34,* 1–100.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In D. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A festschrift for Heinz Neudecker* (pp. 233–247). Dordrecht, Netherlands: Kluwer Academic.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

Tisak, J., & Meredith, W. (1990). Longitudinal factor analysis. In A. von Eye (Ed.), *Statistical methods in longitudinal research: Vol. I. Principles and structuring* (pp. 125–149). New York: Academic Press.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59,* 197–210.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 61–100). Mahwah, NJ: Lawrence Erlbaum Associates.

Webster, H., & Bereiter, C. (1963). The reliability of changes measured by mental test scores. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 39–59). Madison, WI: University of Wisconsin Press.

Wrensen, L. B., & Biderman, M. D. (2005, April). *Factors related to faking ability: A structural equation model application.* Paper presented at the 20th annual Society for Industrial and Organizational Psychology Conference, Los Angeles, CA.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20,* 71–87.

Zickar, M. J., & Gibby, R. E. (2006). A history of faking and socially desirable responding on personality tests. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 21–42). Greenwich, CT: Information Age.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84,* 551–563.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3: Item analysis and test scoring with binary logistic models.* Chicago: Scientific Software.

Controlling social desirability may attenuate faking effects: a study with aggression measures.

Controlar la deseabilidad social puede atenuar los efectos del falseamiento: un Estudio con medidas de agresividad.

Abstract

Background: Several studies have been conducted to better understand what happens with personality scores when faking occurs, but very few use socially undesirable trait measures such as aggression. The aim of the present research is twofold. On the one hand, we aim to apply the procedure by Ferrando, Lorenzo-Seva and Chico (2009) to aggression scales and determine whether it can correct for faking effects. On the other hand we aim to test the impact that individual differences can have on change scores due to faking.

Method: Participants were 371 undergraduate students. Of these, 215 answered the questionnaires twice, under neutral conditions and under faking-motivating conditions. The remaining 156 were the control group who answered the questionnaires twice, both times under neutral conditions.

Results and conclusions: The results showed that the procedure does correct for faking effects and that individual differences have an important impact on the change scores due to faking, except in the most undesirable Physical aggression measure, which was hardly affected.

Keywords: Physical aggression, Verbal aggression, Indirect Aggression, Social desirability, Faking.

Resumen

Antecedentes: Varios estudios han intentado entender qué sucede con las puntuaciones de las medidas de personalidad cuando se produce falseamiento, pero pocos han utilizado medidas socialmente indeseables como la agresividad. El presente estudio tiene dos objetivos principales. Por un lado, se quiere aplicar el método de Ferrando, Lorenzo-Seva y Chico (2009) a las escalas de agresividad para determinar si el método puede corregir los efectos del falseamiento. Por otro, se quiere comprobar el impacto que tienen las diferencias individuales en las puntuaciones de cambio debido a falseamiento.

Método: 371 estudiantes universitarios participaron en el estudio. De ellos, 215 respondieron el cuestionario dos veces, bajo condiciones neutras y bajo condición de falseamiento. Los demás 156 participantes formaron el grupo control contestando dos veces el cuestionario, ambas bajo condiciones neutras.

Resultados y conclusiones: Los resultados muestran que el método corrige los efectos del falseamiento y que las diferencias individuales tienen un papel importante en las puntuaciones de cambio debido a falseamiento, con la excepción de la medida de agresividad más indeseable, la agresividad física, que casi no se ve afectada.

Palabras Clave: Agresividad Física, Agresividad Verbal, Agresividad Indirecta, Deseabilidad Social, Falseamiento

Although personality questionnaires are widely used, they are far from being a perfect measure of the trait/s they intend to measure. In addition to the problem of estimating the true trait level, another important problem is that some subjects might intentionally distort their answers in order to give a better or worse image, especially when they are under pressure and trying to create a positive or negative impression which they believe will increase their chance of achieving certain goals (e.g., Furnham, 1986, Griffith & Peterson, 2008, McFarland & Ryan, 2000).This behaviour is known as faking.

Social Desirability Scales (SDS) were developed to capture a response style used by respondents that is intended to make them appear more favourable than they really are (Paulhus, 1991). Most SDS measure this tendency by using items that are difficult to endorse in a normative sample (Burns & Christiansen, 2006). Thus, SDS measure the respondent's tendency respond to the socially desirable content of the item instead of its trait content (Kuncel, Borneman, & Kiger, 2012). For this reason, SDS are often used as indicators of faking.

As faking-related measures, some authors (e.g. Furnham, 1986, Eysenck & Eysenck, 1976) have interpreted Social Desirability (SD) scores in two ways. When administered under faking-motivating conditions these scores are thought to behave as detection measures because they are highly sensitive to faking. When administered under neutral conditions, however, they are thought to measure a substantive personality variable that has a certain degree of consistency across time and situations (e.g. Furnham, 1986, McFarland & Ryan, 2000).

In many selection and assessment settings, various strategies are used that combine SDS in conjunction with a personality test in order to obtain a more accurate trait estimate under faking-motivating conditions. The earliest technique uses SDS to eliminate

candidates with scores over a certain cut-off point. The two main concerns when using this technique are that, although extreme respondents are removed, the participants who are not cannot be said to be free of SD. Furthermore, SD may be related to such traits of interest as conscientiousness or responsibility and, therefore, deleting the candidates with high SD scores usually involves deleting the candidates who have extreme scores on these traits (McCrae & Costa, 1983; Smith & Ellingson, 2002).

Another commonly used method is partialing or correcting questionnaire scores using SDS. In fact, SDS were originally used to remove the effects of faking by regressing the SD scores onto trait scales and computing a residual score (Meehl & Hataway, 1946). Using correction techniques under neutral and selection conditions, Christiansen, Goffin, Johnston, & Rothstein (1994) found that 70% of the sample was affected by corrections based on SDS, and the rank order changed for 85% of candidates. Thus, decisions based on corrected or uncorrected scores would be markedly discrepant. Nevertheless, various studies have shown that correcting or partialing SD decreases validity: i.e., the partialling of variance associated with SDS may remove meaningful variance from the relevant trait and may decrease the validity of the measures (Li & Bagger, 2006; McCrae & Costa, 1983; Ones, Viswesvaran, & Reiss, 1996; Soubelet & Salthouse, 2011). Furthermore, partialing or correcting also assumes that all items are parallel measures of the trait and this is almost never true (Leite & Cooper, 2010).

Recently Ferrando, Lorenzo-Seva and Chico (2009) proposed a general factor-analytic procedure for assessing response bias in questionnaire measures which may be useful in developing a third approach that overrides the limitations of the previous approaches. The procedure has two main steps. The first step identifies a factor related to SD. To this end a set of items related to SD are selected. These items are known as markers.

The inter-marker correlation matrix obtained is factor analyzed and the corresponding loading values of each marker on the SD factor are estimated. These loading values are then used to compute the loading values of the content items on the SD factor using an instrumental-variables approach (Hägglund, 1982), and the variance explained by the SD factor is removed from the inter-item correlation matrix. In the second step, the residual inter-item correlation matrix is factor analyzed to identify the content factor or factors of interest which are orthogonal to the SD factor.

The application of this procedure at the item calibration level provides two loading estimates for each item: a loading on the content factor that the test wants to measure, and a loading on an orthogonal factor identified as SD. Thus, SD-free content scores are obtained and there is no need to a) assume that items are parallel measurements (which they never are); b) include SDS in the content scales of interest, which considerably increases the questionnaire's length; or c) have a non-faked measure for purposes of comparison, which is practically impossible.

Because scores on both the SD marker items and the content items with high loadings on SD are expected to be more prone to change under faking instructions (Furnham, 1986, Eysenck & Eysenck, 1976), our hypothesis is that under these conditions the SD correction on the content scores will be stronger than under neutral conditions. This is expected to remove (ideally), or at least attenuate, the effects of faking on the content scores.

When faking occurs, it is important to know the extent to which individual differences affect the magnitude of the change in the scores due to similar faking conditions. However, recent reviews (Burns & Christiansen, 2006; Mesmer-Magnus & Viswesvaran, 2006) show that there is very little literature on this issue. What is of most interest is to

investigate whether the magnitude and the direction of the change is the same for all subjects or whether the change is specific to every single subject. If all subjects change in exactly the same way, individual differences have no effect on the amount of change, the rank order is not affected by the faking instructions and, therefore, controlling the amount of change would make no difference in selection. On the other hand, if individual differences impact the amount of faking-related change, those subjects who modify their scores in the most appropriate direction will have an unfair advantage over the honest subjects. Ferrando and Anguiano-Carrasco (2011) assessed this issue and found that individual differences have an important impact on the Psychoticism and Neuroticism scales of the Eysenck Personality Questionnaire.

Unlike previous research, the present research uses measures of such highly undesirable behaviours as aggression. Faking is expected to have greater effects on these measures as individuals want to give a good impression. As for the impact of SD, the research generally shows a moderate-to-high relationship between SD and aggression measures. Biaggio (1980) and Selby (1984) reported that most of the correlations between the Buss-Durkee Hostility Inventory scales and the Marlowe-Crowne Social Desirability scale were in the range r = -.3 to -.5. SD has also been related to measures of violent behaviours and partner abuse (Bell & Naugle, 2007; Devon, Collie & Walkley, 2004) and those aspects of NEO-PI-R scales most related to aggressive behaviour such as impulsivity and angry hostility (Holden & Passey, 2010). Recently Vigil-Colet, Ruiz-Pàmies, Anguiano-Carrasco and Lorenzo-Seva (2012) used the same method as the one used in the present research in a study based on neutral conditions. Results showed (a) that the items on the aggression questionnaires have moderate-to-high loadings on the SD factor, and (b) that when corrected for this effect, the scores on the aggression scales

tended to increase considerably. Conceptually these results suggest that (a) the chosen measures are clearly impacted by SD and (b) the method corrects in the expected direction. Consequently, they are the basis for the present research, which can essentially be considered as an extension of the study by Vigil-Colet et al. (2012) in which the scores are obtained under both neutral and faking-inducing conditions.

The two aggression measures used in Vigil et al. (2012) were: (a) Buss and Perry Aggression Questionnaire (BPAQ; Buss & Perry, 1992) which has proved to be useful in assessing various levels and types of direct aggression (e.g. Morales-Vives & Vigil-Colet, 2010), and (b) the Indirect Aggression Scale (IAS; Forrest, Eatough & Shevlin, 2005). The BPAQ is intended to measure four aggression scales: Physical aggression, Verbal aggression, Anger and Hostility. However, the factorial structure of the BPAQ remains controversial, generally due to the scales intended to measure anger and hostility. As for the IAS, it was included because indirect aggression (see Björkqvist, Osterman & Kaukiainen, 1992), which has been shown to be the most usual type of aggression in adults, is not considered in the BPAQ.

Overall, the present research used the Physical, Verbal (BPAQ) and Indirect (IAS) aggression scales. The Anger and Hostility scales of the BPAQ were avoided for two main reasons. On the one hand, the procedure for "cleaning" the content scores of SD uses the residual correlation matrix, so using dimensions which often present unstable solutions would only produce confounding and unsettled results. On the other hand, Anger and Hostility are defined as feelings and cognitions that are strongly related to aggression but they cannot be considered to be aggressive behaviour.

To assess our main hypothesis we proposed a repeated measures design with two factors: condition (neutral vs faking) and correction (with or without the proposed SD

correction). Our hypothesis is that if the proposed correction reduces faking effects, then we will find an interaction between condition and correction in the sense that the content scores are less affected by faking under the correction condition. To assess the second important issue in the present research, – the impact of individual differences on change scores due to faking, – we used the same procedure and statistics as the ones described by Ferrando and Anguiano-Carrasco (2011).

Method

Participants

Participants were 371 undergraduate students from different faculties of the Rovira i Virgili University (Spain). They were randomly assigned in class groups to experimental or control groups. The control group was made up of 156 students and the experimental group of 215 students. The groups were comparable: 85% were women and the mean age was 21 years old in both. The questionnaires were administered in paper and pencil version by the same person in all cases, and completed voluntarily in classroom groups of 25 to 60 students. The administration was anonymous, and the respondents had to provide only three particulars which were used for matching: gender, date of birth and favourite colour.

Procedure

All participants filled in the questionnaires twice. The participants in the control group were asked to respond twice under the standard instructions provided in questionnaires. Among other things, the instructions advise participants to give honest answers. The participants assigned to the experimental group were divided into two subgroups, one of

which was first given the faking-motivating instructions and then, on the retest, asked to respond honestly. The other half was first instructed to answer honestly and then, on the retest, given the faking-motivating instructions. The faking-motivating instructions were those listed in Eysenck, Eysenck and Shaw (1974). Respondents are asked to imagine that they are applying for a job that they really want. They should try to give a good impression by answering what they think the employer would like to hear. The re-test interval was six weeks in all cases.

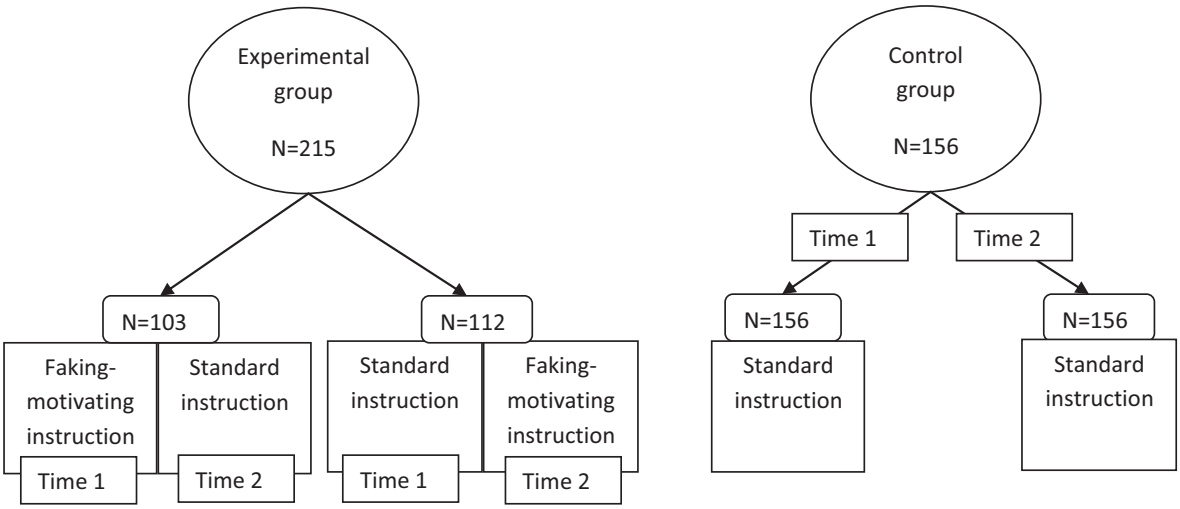


Figure 1. *Experimental design: groups and instructions given.*

Instruments

The study used the Physical and Verbal aggression scales (7 and 4 items respectively) of the Spanish short version of the BPAQ (Vigil-Colet, Lorenzo-Seva, Codorniu-Raga & Morales, 2005) as well as the Indirect Aggression Scale (IAS), in the Spanish short version (10 items) by Anguiano-Carrasco & Vigil-Colet (2011). Overall, given that the

procedures in Ferrando et al. (2009) provide content and SD scores for each measure, the analyses that follow are based on five sets of individual scores: Physical aggression, Verbal aggression, Indirect aggression, BPAQ SD and IAS SD.

Data Analysis

Only the experimental group was used to assess the first issue. To assess the second issue we used a structural equation model (SEM) in which the amount of individual change is estimated on the basis of a bidimensional invariant model (Ferrando & Anguiano-Carrasco, 2011). Both control and experimental groups were used in the second assessment.

To assess the first issue, we obtained the effects of the SD corrections under neutral and faking conditions, the content and the SD factor scores using the following procedure. First, we used separate factor analyses and checked that the loading estimates obtained under the neutral and faking conditions were essentially invariant (they were). Second, common estimates were obtained by averaging the loadings obtained in both conditions. Third, Bartlett factor scores (see e.g. Ferrando 2007) were obtained based on these common estimates. In the "corrected" conditions, the factor-score estimates were based on the bidimensional (content and SD) solution of Ferrando et al.. In the non-corrected conditions they were based on the undimensional solution. The scores were standardized in the complete dataset, so that results were comparable when the dataset was split into the different conditions.

Results

Table 1 shows the mean scores on the T scale for the four conditions for each of the three scales used in the study and for SD. The table shows that all the scales corrected

by the procedure have higher means than the uncorrected ones, as expected given that higher scores imply higher levels of aggression. The neutral scores are also higher than the faked ones for each content scale. The SD scores were expected to be sensitive to faking, and higher in the faking condition. We found that the scores for the two SD measures (SD computed on BPAQ and IAS) were significantly greater (t = 14.53, p = 0.01; t = 13.27, p = 0.01; respectively) with effect sizes of d = 1.17 and d=1.06. According to Cohen's criteria (1969, p. 23), these may be considered to be large.

Table 1. Mean T scores for each scale on each condition.

	Non-Corrected		**Corrected**	
	Neutral	**Faked**	**Neutral**	**Faked**
Physical aggression	44.54 (5.52)	41.56 (3.47)	52.43 (5.35)	51.79 (3.40)
Verbal aggression	48.38 (8.72)	39.88 (8.56)	49.32 (8.64)	42.54 (8.43)
Indirect aggression	47.47 (9.96)	40.60 (6.56)	57.62 (9.07)	52.07 (5.91)
SD on BPAQ			66.23 (6.04)	75.43 (5.62)
SD on IAS			67.62 (7.57)	76.10 (5.46)

Note. In Brackets Standard deviation.

Table 2 shows the results of the two by two (corrected vs. uncorrected, neutral vs. faked) factor repeated measures analysis of variance for each scale. Both factors and their interaction showed significant effects on all scales, so the score changes related to faking depended upon the presence or absence of SD correction. The partial Eta squared statistic, is also shown. Figure 2, shows the interaction effects on each scale. As can be seen, the differences between the scores under faking and neutral conditions are always smaller for corrected scores, and are even non-significant in the case of physical aggression (t = -1.56, p = 0.119). For verbal and indirect aggression, on the other hand, there is a reduction in faking effects but the difference between both conditions is still significant (t = 7.76, p = 0.01; t = 6.90, p = 0.01, respectively).

Table 2. Univariate contrast. F statistic, its significance level and partial Eta Squared for principal factors and their interaction for each scale.

		F	***P***	η_p^2
Physical aggression	**Answer Condition**	31.57	0.00	0.06
	Correction Condition	5452.54	0.00	0.80
	Interaction C × C	125.63	0.00	0.45
Verbal Aggression	**Answer Condition**	112.02	0.00	0.32
	Correction Condition	580.27	0.00	0.88
	Interaction C × C	187.88	0.00	0.47
Indirect Aggression	**Answer Condition**	191.40	0.00	0.25
	Correction Condition	245.96	0.00	0.94
	Interaction C × C	80.65	0.00	0.32

Note. F = F statistic; P = F's probability; η_p^2 = partial squared eta.

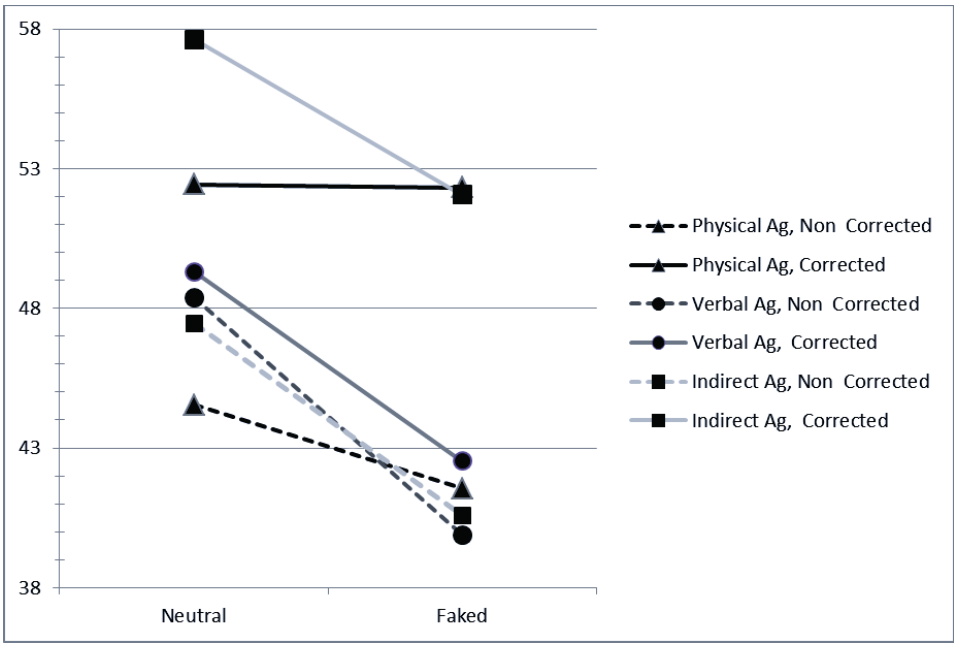


Figure 2. *Interaction effects on the aggression scales.*

Table 3 shows the correlations between the increments in SD and in the aggression scale scores when subtracting the T scores under faking conditions from the T scores under neutral conditions. All the correlations except those with the Indirect Aggression Scale were significant, showing that in direct aggression measures the change in SD items is related to the change in content measures.

Table 3. Correlations between SD score increment and the increments on each aggression scale.

	SD
Physical aggression	0.36**
Verbal aggression	0.33**
Indirect aggression	-0.11

Note. ** significance level 0.01.

In order to test the hypothesis that individual differences play an important role in scale change scores due to faking, the three scales were compared by fitting them on the bidimensional invariant and non-invariant models. The invariant model indicates that the factor under neutral conditions has exactly the same structure, factor loadings and thresholds as under faking-inducing conditions so the scores obtained under neutral and faking conditions can be compared. Table 4 shows that the fit of the invariant model was acceptable for all the scales and not substantially worse than the fit of the less restrictive non-invariant model. We therefore consider the invariant model to be acceptable.

Table 4. Goodness of fit statistics for invariant and non-invariant models of P.A., V.A. and I.A.

	χ^2		Df		CFI		RMSEA	
	Non-invariant	Invariant	Non-invariant	Invariant	Non-invariant	Invariant	Non-invariant	Invariant
P.A. control	64.48	99.80	47	57	0.96	0.92	0.04	0.06
P.A. faked	71.81	85.89	47	57	0.93	0.92	0.05	0.05
V.A. control	9.71	19.83	5	10	0.97	0.95	0.07	0.07
V.A. faked	6.41	12.93	5	10	0.99	0.98	0.03	0.03
I.A. control	151.74	196.11	125	142	0.96	0.93	0.03	0.05
I.A. faked	174.30	237.27	125	142	0.95	0.92	0.04	0.05

Note. P.A.= Physical aggression, V.A. =Verbal aggression and I.A. = Indirect aggression.

We next estimated the amount of relative variance to assess the impact of individual differences (measured by Cohen's d) on the amount of change caused by faking-inducing instructions for each scale. Those results showed that the most impacted measure was Indirect aggression (d = 2.95), followed by Verbal aggression (d = 2.24) and then Physical aggression (d = 0.65).

Discussion

The aim of the study was twofold. First, it aimed to confirm the hypothesis that the procedure proposed by Ferrando, et al. (2009) reduces the faking effect, and that the SD

factor obtained is highly affected by faking-induced change and may be useful for correcting the scores on the scales that change the most under faking inducing instruction. Second, we assessed the impact that individual differences have on the change scores due to faking on the aggression measures. We were particularly interested in determining whether the scales that are most impacted by individual differences are also the ones in which increments in SD do not correlate with the increments in the scale scores.

The results suggests that, although far from being perfect, the procedure is useful for 'cleaning' the scores and attenuating the effects of faking-inducing instructions because it has a differential effect on the increments caused by faking on the aggression measures. Cohen's d indicates that the SD is very sensitive to faking-inducing instructions and it has a big effect on both questionnaires. The correlations show that in direct aggression measures the amount of change due to faking is related to the increments in the SD scores that the procedure provides.

The results also seem to indicate that SD factor increments are clearly related to increments due to faking-inducing instructions on the aggression scales with the exception of Indirect aggression. This result could be explained by the fact that Indirect aggression is the most acceptable, socialized type of aggression, so when subjects fake they do not consistently change their scores on Indirect aggression in the same direction or magnitude. This conjecture is supported by the result that individual differences have the biggest impact on this scale.

Individual differences explain quite a large amount of the total variance in verbal and indirect aggression, but not so much of the variance in physical aggression. However, although it is clear that individual differences have less impact on physical aggression,

according to Cohen's criterion their effect would still be medium. We should point out here that the physical aggression scale showed the smallest overall variance. In our opinion the fact that the overall variance is small is one reason why the relative importance of the individual differences variance appears to have a medium effect size although the direct measure is not very big. Therefore, we consider here that individual differences have a very small, almost negligible, impact.

Physical aggression is considered to be the predominant type of aggression in children but it progressively decreases during the socialization process. Verbal and Indirect aggression become more important and peak during adolescence and adulthood (Vaillancourt, 2005; Tremblay & Nagin, 2005). It is, therefore, reasonable to suggest that physical aggression is the most socially undesirable behaviour of all the aggression types assessed in the present research. Taking into account everything explained above, we conjecture that the impact of individual differences on highly undesirable behaviours is negligible in terms of rank order: that is, all the subjects increase or decrease their scores by the about same magnitude and in the same direction. Therefore, rank order in a possible personnel selection, or any situation in which the extreme scoring subjects are to be selected, would not be affected by faking on these types of measure. As can be seen, physical aggression is almost not impacted by individual differences in faking change scores but verbal and indirect aggression, which are more acceptable aggression behaviours, are.

Consequently, it would be of interest to measure how different personality traits are affected by individual differences in faking. If the results obtained here are generalizable to other behaviours considered to be extremely undesirable, the decisions

based upon individuals' scores may be correct even though they may be affected by faking.

No study is free of limitations and the present one is no exception. On the one hand, our participants were university students instructed to fake, not real job applicants or patients. It would be desirable to compare the results obtained here with results from samples of real job applicants or patients. On the other hand, in order to consolidate the procedure and generalize its use it would be of interest to replicate the results of this research on such trait scales as Conscientiousness or Integrity, which have proved to be closely related to SD (McFarland & Ryan, 2000; Muller-Hanson, Heggestad, & Thornton, 2006; Griffith, Malm, English, Yoshita, & Gujar, 2006).

In conclusion, the factor analytic procedure proposed by Ferrando et al. appears to be an important tool for controlling the effect that faking has on personality scale scores. The procedure only needs the four selected markers to be added to the scale of interest and to be administered once. The test, then, is not excessively longer and there is no need for initial scores to be neutral, which is by no means easy to achieve in such contexts as clinical assessments or personnel selection procedures.

Acknowledgements

The research was supported by a grant from the Spanish Ministry of Economy and Competitiveness (PSI2011-22683).

References

Anguiano-Carrasco, C., & Vigil-Colet, A. (2011). Assessing indirect aggression in aggressors and targets: Spanish adaptation of Indirect Aggression Scales. *Psicothema, 23,* 146-152.

Bell, K.M.,& Naugle, A.E. (2007). Effects of social desirability on students' self-reporting of partner abuse perpetration and victimization. *Violence and Victims, 22*, 243-256.

Biaggio, M.K. (1980). Assessment of anger arousal. *Journal of Personality Assessment, 44,* 289-298.

Björkqvist, K., Osterman, K., & Kaukiainen A. (1992). The development of direct and indirect strategies in males and females. In K.Bjorkqvist, and P.Niemela(Eds.) *Of mice and women: Aspects of female aggression (pp.51-64).* San Diego, CA: Academic Press.

Burns, G.N., & Christiansen, N.D. (2006). Sensitive or senseless: On the use of social desirability measures in selection and assessment. In R.L. Griffith & M.H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 113-148). Greenwich, CT: Information Age Publishing.

Buss, A. H., & Perry, M. P. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology, 63,* 452-459.

Christiansen, N.D., Goffin, R.D., Johnston, N.G., & Rothstein, M.G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47,* 847-860.

Devon, L. L., Collie, R.M., & Walkley, F.H. (2004).Criminal attitudes to violence: Development and preliminary validation of a scale for male prisoners. *Aggressive Behavior, 30*, 484-503.

Eysenck, H.J. & Eysenck, S.B.G. (1976). *Psychoticism as a dimension of personality*. New York: Crane-Russak.

Eysenck, S.B.G., Eysenck, H.J., & Shaw, L. (1974). The modification of personality and Lie scores by special 'honestly' instructions. *British Journal of Social and Clinical Psychology,* 13, 41-50.

Ferrando, P.J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research, 42*, 481-508.

Ferrando, P.J., & Anguiano-Carrasco, C. (2011). A Structural Equation Model at the Individual and Group Level for assessing faking-Related Change. *Structural Equating Modeling, 18,* 91-109.

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.

Forrest, S., Eatough, V. & Shevlin, M. (2005) .Measuring adult indirect aggression: The development and psychometric assessment of the indirect aggression scales. *Aggressive Behavior,31,* 84-97.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, 7, 385-400.

Griffith, R.L., & Peterson, M.H. (2008). The Failure of Social Desirability Measures to Capture Applicant Faking Behavior. *Industrial and Organizational Psychology, 1,* 308-311.

Griffith, R.L., Malm, T., English, A., Yoshita, Y., & Guajar, A. (2006). Applicant faking behavior: Teasing apart the influence of situational variance, cognitive bias, and individual differences. In R.L. Griffith & M.H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 151-178). Greenwich, CT: Information Age Publishing.

Hägglund, G. (1982). Factor analysis by instrumental variable methods. *Psychometrika, 47,* 209–222.

Holden, R. R., & Passey, J. (2010). Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and individual differences, 49,* 446-450.

Kuncel, N.R., Borneman, M., & Kiger, T. (2012). Inovative Item Response Process and Bayesian Faking Detection Methods. . In M. Ziegler, C. MacCann & R.D. Roberts (Eds.), *New Perspectives on Faking in Personality Assessment* (pp. 72-84). NY: Oxford University Press.

Leite, W. L., & Cooper, L. A.(2010). Detecting Social Desirability Bias Using Factor Mixture Models. *Multivariate Behavioral Research, 45*, 271- 293.

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection & Assessment, 14,* 131-141.

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: more substance than style. *Journal of Consulting and Clinical Psychology*, *51*, 882–888.

McFarland, L.A., & Ryan, A.M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology,* 85, 812-821.

McFarland, L.A., & Ryan, A.M. (2006). Toward and Integrated Model of Applicant Faking Behavior. *Journal of Applied Social Psychology*, 36, 979-1016.

Meehl, P.E., & Hathaway, S.R. (1946). The K factor as a suppressor variable in the MMPI. *Journal of Applied Psychology, 30,* 526-564.

Mesmer-Magnus, J., & Viswesvaran, C. (2006). Assessing response distortion in personality tests: A review of research designs and analytic strategies. In R.L. Griffith & M.H. Peterson (Eds.), *A closer examination of applicant faking behaviour* (pp. 85-113). Greenwich, CT: Information Age Publishing.

Morales-Vives, F., & Vigil-Colet, A. (2010). Are there sex differences in physical aggression in the elderly?. *Personality and Individual Differences, 49,* 659–662.

Mueller-Hanson, R.A., Heggestad, E.D., & Thorton, G.C. (2006). Individual differences in impression management: an exploration of the psychological process underlying faking. *Psychology Science,* 48, 288-312.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*, 660–679.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds*.), Measures of personality and social psychological attitudes* (pp. 17–59). San Diego: Academic Press.

Selby, M.J. (1984). Assessment of violence potential using measures of anger, hostility, and social desirability. *Journal of Personality Assessment, 48,* 531-544.

Smith, D. B., & Ellingson, J.E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology, 87*, 211-219.

Soubelet, A., & Salthouse, T.A. (2011). Influence of Social Desirability on Age Differences in Self-Reports of Mood and Personality. *Journal of Personality, 79*, 741-762.

Tremblay, R.E., & Nagin, D.S. (2005). The Developmental Origins of Physical Aggression in Humans. En R. E. Tremblay., W.W. Hartup., y J.Archer (Eds.) *Developmental origins of aggression (pp.83-106)*. New York. Guilford Press.

Vaillancourt, T. (2005). Indirect aggression among humans: social construct or evolutionary adaptation?. En R. E. Tremblay., W.W. Hartup., y J.Archer (Eds.) *Developmental origins of aggression (pp.158-177)*. New York. Guilford Press.

Vigil-Colet, A., Lorenzo-Seva, U., Codorniu-Raga, M.J., & Morales, F. (2005). Factor structure of the aggression questionnaire among different samples and languages. *Aggressive Behavior, 31,* 601–608.

Vigil-Colet, Ruiz-Pàmies, Anguiano-Carrasco, & Lorenzo-Seva. (2012). The impact of social desirability on psychometric measures of aggression. *Psicothema,24,* 310-315.

4. Discussion

4. Discussion

To achieve our first goal, we adapted the Indirect Aggression Scale by Forrest, Eatough, & Shevlin (2005). This scale was originally developed to measure three facets of indirect aggression in the general life of adults, not in work-related situations. These factors were social exclusion, guilt induction and malicious humor. This instrument has two main properties that made it our choice for adaptation. First, it measures only indirect aggression; the items are not a mix of direct and indirect aggression, as in other questionnaires. Second, it has two forms: one for aggressors and one for targets. The items are essentially the same but change only the focus of the aggressive behavior: you in the case of the target form (IAS-t) and others in the case of the aggressor form (IAS-a). For the purposes of the present study, only the aggressor form was required, but we thought it was of interest for applied psychologist to have both forms available. The Kaiser-Mayer-Olkin index showed that the amount of inter-item consistency was appropriate for carrying out factor analysis: it was 0.91 for IAS-a and 0.93 for IAS-t. In our case both parallel analyses and the scree test agreed that one dimension was underlying each form.

All items showed appropriate discriminating power, with standardized factor loadings, greater than 0.40, and item-total correlations between 0.30 and 0.60. Thus, no item needed to be deleted and the adapted forms were finally made up of 25 items each. Taking into account the good psychometric properties shown by the items we decided that a shorter scale could be developed for each version. We chose the ten items with the highest factor loadings on each scale to create the short versions of the questionnaires.

Finally we examined the properties of the full- and short-version scales. No significant sex differences were found for any of the scales in either the full or the short versions. Reliabilities were also estimated using Cronbach's alpha statistic. For all scales, reliabilities were high and ranged from α =.818 to α =.898. Short versions did not significantly decrease the reliability of the scales, so they can be used without any loss in reliability. The IAS-a scale also showed good convergent and criterion validity as indicated by the product moment correlations between the scales and other aggression scales (BPAQ) and with dysfunctional impulsivity (DII). On the other hand, IAS-t had no correlation with other aggression scales or with impulsivity, as

expected, with the exception of the hostility scale. Taking all this into account, we can conclude that the Spanish full and short versions of IAS are one-dimensional scales that present good reliabilities and validities.

With the exception of the factor structure, our results are in the same direction as the ones obtained by the original authors. Although the original version by Forrest et al. found a three-dimensional structure for each scale, it had some methodological limitations that may raise doubts about the results. First, the number of factors was determined using the Kaiser rule (1970), which tends to overestimate. Second, the Pearson correlation matrix was used to extract the factor loadings, when the polychoric correlation matrix is possibly more appropriate in this case (because of the relatively high item discriminations). Third, an orthogonal rotation procedure was used, but it is hard to assume that different forms of indirect aggression are independent. Warren, Richardson and McQuillin (2011) examined the nature of indirect aggression using different indirect aggression questionnaires that measured different facets of the trait. They found that the questionnaires overlap and deduced that they probably measure the

same basic construct. Their findings support the factor structure shown in our study and the one-dimensionality of indirect aggression.

Our second goal was to develop new psychometric procedures that allow the researcher to estimate the amount of variance in change scores that is accounted for by faking. On the basis of the theta-shift model (Zickar & Drasgow, 1996; Zickar & Robie, 1999), we assumed that under faking motivating conditions subjects temporary change their true trait level. This approach overcomes some of the limitations that are present when working with raw change scores, such as end effects or measurement error, by working at the theta level. We assessed the model at group and individual level.

At the group level the amount of change due to faking-inducing instructions provides the researcher with information about how individual differences impact on change scores and shed some light on faking behavior. When we used an Impression Management (IM) scale, which has proven to be very sensitive to faking, results at the group level indicate that participants who have already put themselves in a favorable light under neutral answering instructions tend to change their scores very little when asked to fake good. On the other

hand, participants who showed initial low IM scores under neural answering conditions tended to change their scores considerably when asked to fake.

At the individual level, the inconsistency of the individual should first be assessed so those participants who are not consistent can be deleted. Subsequently, the researcher can estimate the amount of variance due to the faking-inducing instructions in change scores for each individual. The statistic obtained should give some idea of the effective faking that the individual has applied so that decisions about the trustability of each participant's scores can be taken. In summary, we have provided the researcher with a detection tool that can be useful in some applied areas (for example, when evaluating an inmate for prison leave). But it is also true that we do not have a cut-off value, so there is no standard point from which the researcher or the applied psychologist can affirm that the respondent has faked. However, the statistic gives some extra information about the respondent's behavior during the assessment process, and he/she can take their own decision about how much faking he/she is willing to take.

Finally we shall discuss our last goal. First we focus on the effects that faking has on highly undesirable personality traits, as is the case for aggression. We used the correction method by Ferrando, Lorenzo-Seva and Chico (2009) to examine how raw scores change when corrected for faking effects and how individual differences affect the change scores due to faking. We also take the opportunity to examine if the correction method indeed corrects in the expected direction, and to check that correcting for social desirability will in fact correct for faking. We found that aggression scales, as measures of an undesirable trait, are corrected to a considerable extent by the method, which supports the results obtained by Vigil-Colet et al. (2012). We also have evidence to suggest that the correction effects are in the expected direction. Thus, although we cannot say that the faking effects are totally removed from the trait scales scores, they are mitigated. We also examined the role that individual differences have on the change scores due to faking-inducing instructions. Previous research using different personality scales showed that individual differences had quite an important role in the change scores (Ferrando & Anguiano-Carrasco, 2011b; Ferrando & Anguiano-Carrasco, 2011c), meaning that each individual has his/her own idea about what socially

accepted behavior is or the extent to which it is “allowed” to fake, and they modify their scores accordingly. This is also the case for indirect and verbal aggression, but not for physical aggression. Physical aggression is the most undesirable personality trait of the ones measured so far, and in our society, it is clearly reprehensible behavior, usually condemned by others and even prosecuted. In contrast, indirect aggression is the most accepted kind of aggression. It has been socialized, and although it is not encouraged, it is not considered to be totally reprehensible. Gossiping, for example, can be considered as a perfectly normal passtime as can be seen in the television schedule every day. Verbal aggression is also considered to be quite an acceptable kind of aggression, and is socially permitted (at least in our culture). Raising your voice when arguing, for example, is normal behavior (in Spain) and not perceived as an act of aggression (at least not by everybody). Once again television provides examples of this behavior in everyday gatherings. No matter what topic is being debated, verbal aggression is often triggered and is not considered as a behavior that should be avoided. In some cases it is even encouraged by moderators and audience. Thus, indirect aggression and verbal aggression is deeply impacted by individual differences, as

shown by Cohen's d statistic in the third paper presented in this study. However, physical aggression, a behavior widely recognized as undesirable and reprehensible, is not.

The field of response distortion of personality traits is still widely unknown and faking is one of the important types of response distortion on which researchers have focused. Although there is a long tradition of studying faking, very little is known and numerous questions remain unsolved. The present study aims to shed light on the field but its contribution is admittedly limited. On the one hand, we have developed a method that allows the researcher to explore change scores under faking-motivating conditions at the trait level, overcoming some of the problems that arise when working with raw scores, although the conditions needed for this method to be implemented are hard to achieve. It is certainly very difficult for researchers to get neutral scores on a test, even if no specific instructions are given or the situation does not encourage respondents to fake. People tend to give a better image of themselves when answering a questionnaire, especially if it deals with highly

undesirable personality traits as is the case of aggression measures. As the third paper presented in this dissertation shows, even if participants who are inconsistent in their neutral condition scores are removed, faking still occurs as their scores change substantially when corrected using the Ferrando, Lorenzo-Seva and Chico (2009) factor analytic procedure. Even if it is assumed that respondents will be honest when they answer, it is difficult to find a natural situation in which a researcher or an applied psychologist will assess participants twice, once under neutral conditions and once under faking-motivating conditions. Maybe the case of prison leaves is the clearest one, but it is not very common, so its applicability is quite limited. From a research point of view the impact that individual differences have on change scores due to faking is an interesting new field that sheds light on how faking behaves. Various studies (e.g. Ferrando & Anguiano-Carrasco, 2011b; Ferrando & Anguiano-Carrasco, 2011c; Anguiano-Carrasco, Vigil-Colet & Ferrando, in press) have shown that individual differences do not impact to the same extent on all personality measures, but the nature of the trait studied also has an important role.

Nevertheless, aggression is still a personality trait that must be studied in depth if it is to be understood more fully. Specifically, the role of anger and hostility is not clear and, as far as we know, no questionnaires appropriately measure direct and indirect aggression at the same time. The lack of appropriate instruments is even more acute if we aim to control for response bias. The adaptation of an indirect aggression scale was the first attempt to obtain a reliable measure for indirect aggression (Anguiano-Carrasco & Vigil-Colet, 2011) but biases were not directly controlled and a second scale for direct aggression should be used if all types of aggression are of interest. These issues should be investigated in further research in order to better assess a personality trait that is becoming more and more important in our society.

To summarize, the present dissertation brings together most of the existing methods of controlling and assessing faking and proposes a new method for assessing how faking impacts on personality measures. It also proposes an adaptation of an indirect aggression scale and applies the new method not only to this questionnaire but

also to the Verbal and Physical Aggression scales of BPAQ. It also assesses whether the factor-analytic correction method proposed by Ferrando, Lorenzo-Seva and Chico (2009) in fact corrects for faking and examines the role of individual differences in such undesirable personality measures.

To conclude we would like to say that a considerable amount of research still has to be done in the field of faking, as none of the methods examined here has proved to perfectly detect or correct faking, although the structural model-based optimal person fit detection procedure and the factor analytical correction method appear to be the most promising.

4.1. Conclusions

A valid and reliable self-informed questionnaire of indirect aggression was adapted. A short form is also available and it can be used to assess both aggressors and targets.

A new procedure for assessing the amount of trait-level change due to faking was developed. It also accounts for the amount of variance due to individual differences. The method was successfully implemented

on different personality traits. The impact of individual differences on change scores depends on the personality trait studied.

Aggression measures are deeply impacted by faking. Correcting for social desirability actually mitigates the impact of faking on personality scores when using the factor analytic procedure by Ferrando, Lorenzo-Seva and Chico (2009). Although faking effects are not completely controlled, they have been shown to be mitigated by the procedure.

5. References

5. References.

Aguilar, A., Tous, J.M., & Andrés, A. (1990). Adaptación y estudio psicométrico del EPQ-R [Adaptation and psychometric analysis of the EPQ-R]. *Anuario de Psicología, 46,* 101-118.

Anguiano-Carrasco, C., & Vigil-Colet, A. (2011). Assessing indirect aggression in aggressors and targets: Spanish adaptation of the Indirect Aggression Scales. *Psicothema, 23,* 146-152.

Anguiano-Carraso, C., Vigil-Colet, A., & Ferrando, P.J. Controlling social desirability may attenuate faking effects: a study with aggression measures. *Psicothema,* (In press).

Bell, K.M., & Naugle, A.E. (2007). Effects of social desirability on students' self-reporting of partner abuse perpetration and victimization. *Violence and Victims, 22*, 243-256.

Berkowitz, L. (1993). Pain and aggression: some findings and implications. *Motivation and Emotion, 17,* 277-293.

Berkowitz, L. (1994). Is something missing? Some observations prompted by the cognitive-neoassociationist view of anger and emotional aggression. In R. Huesman (Ed.) *Aggressive behavior. Current Perspectives* (pp. 35-60). New York: Plenum.

Berkowitz, L. (1996). Agresssion. Causes, Consequences and control. New York: McGrau Hill

Biaggio, M.K. (1980). Assessment of anger arousal. *Journal of Personality Assessment, 44,* 289-298.

Björkqvist, K. (1994). Sex differences ins physical, verbal and indirect aggression: review of recent research. *Sex Roles, 30,* 177-188.

Björkqvist, K., Osterman, K., & Kaukiainen A. (1992). The development of direct and indirect strategies in males and females. En K.Bjorkqvist, y P.Niemela(Eds.) *Of mice and women: Aspects of female aggression. (pp. 51-64)* San Diego, CA: Academic Press

Borkenau, P., & Ostendorf, F. (1992). Social Desirability as a Moderator and Supresssor Variables. *European Journal of Personality, 6,* 199-214.

Brirkeland, S.A., Manson, T.M., Kisamore, J.L., Brannick, M.T., & Smith, M.A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14,* 317-335.

Brown, R.D. (1997). *The development of a computer adaptive test of the five factor model of personality: Applications and extensions*. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University.

Brown, R.D., & Harvey, R.J. (2003, April). *Detecting personality test faking with appropriateness measurement: Fact or fantasy?* Paper presented at the 2003 Annual Conference of the Society for Industrial and Organizational Psychology, Orlando.

Burns, G.N., & Christiansen, N.D. (2006). Sensitive or senseless: On the use of social desirability meaures in selection and assessment. In R.L. Griffith & M.H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 113-148). Greenwich, CT: Information Age Publishing.

Chico, E., Tous, J.M., Lorenzo-Seva U., & Vigil-Colet A. (2003). Spanish adaptation of Dickman's impulsivity inventory, its relationship to Eysenck's personality questionnaire. *Personality and Individual Differences*, *35,* 1883-1892.

Christiansen, N.D., Goffin, R.D., Johnston, N.G., & Rothstein, M.G. (1994). Correcting the 16PF for faking-effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47,* 847-860.

Christiansen, N.D., Robie, C., & Bly, P.R. (2005, April). Using covariance to detect applicant response distortion of personality measures. Paper presented in M. Zickar (Chair). *Faking research: New methods, new samples, and new questions.* Symposium conducted at the 20th annual conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

Costa, P.T., & McCrae, R.R. (1992). *Revised NEO Personality Inventory and NEO Five Factor Inventory.* Odessa, Fl: Psychological Assessment Resources.

Crowne, D. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349-354.

Drasgow, F., Levine, M.V., & Williams, E.A. (1985). Approproateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Devon, L. L., Collie, R.M., & Walkey, F.H. (2004).Criminal attitudes to violence: Development and preliminary validation of a scale for male prisoners. *Aggressive Behavior, 30*, 484-503.

Dudley, N.M., McFarland, L.A., Goodman, S.A., Hunt, S.T., & Sydell, E.J. (2005). Racial differences in socially desirable responding in selection contexts: Magnitude and consequences. *Journal of Personality Assessment, 85,* 50-64.

Eckhardt, C., & Deffenbacher, J. (1995). Diagnosis of anger desorders. In H. Kassinove (Ed.), *Anger disorders: definition, diagnosis, and treatment* (pp. 27-48). Washington, DC: Taylor & Francis.

Ellingson, J.E., Sackett, P.R., & Hough, L.M. (1999). Social desirability corrections in personality measurement: Issues of applicant

comparison and construct validity. *Journal of Applied Psychology, 84,* 155-166.

Eysenck H. J. & Eysenck S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire.* London: Hodder & Stoughton.

Ferrando, P.J. (2002). An IRT-Based Two-Wave Model for Studying Short-Term Stability in Personality Measuremnt. *Applied Psychological Measurement, 26,* 286-301.

Ferrando, P.J., & Anguiano-Carrasco, C. (2011a). Faking propensity and faking-related change: A model-based analysis of the EPQ-R scores. *Personality and Individual Differences, 51,* 497-501.

Ferrando, P.J., & Anguiano-Carrasco, C. (2011b). A Structural Equation Model at the Individual and Group Level for Assessing Faking-Related Change. *Structural Equation Modeling, 18,* 91-109.

Ferrando, P.J., & Anguiano-Carrasco, C. (2011c). Evaluación de las diferencias individuales en falseamineto. *Psicothema, 23,* 839-844.

Ferrando, P.J., & Anguiano-Carrasco (2012). An structural model-based optimal person fit procedure for identifying faking. *Educational and Psychological Measurement. Under review.*

Ferrando, P.J., & Chico, E. (2001). Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales. *Educational and Psychological Measurement, 61*,997-1012.

Ferrando, P. J., Chico, E., & Lorenzo, U. (1997). Dimensional analysis of the EPQ-R Lie scale with a Spanish sample: Gender differences and relations to N, E, and P. *Personality and Individual Differences, 23,* 631–637.

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.

Forrest, S., Eatough, V., & Shevlin, M. (2005). Measuring adult indirect aggression: The development and psychometric assessment of the indirect aggression scales. *Aggressive Behavior, 31,* 84-97.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7,* 385-400.

Gough, H.G., & Bradley, P. (1996). *CPI manual.* Palo Alto, CA: Consulting Psychological Press.

Hägglund, G. (1982). Factor analysis by instrumental variable methods. *Psychometrika, 47,* 209–222.

Hambleton, H.K., & Swaminathan, H. (1989). Item Response Theory. Principles and Applications. Boston: Kluwer Nijoff Publishers.

Harvey, R.J., Wilson, M.A., & Hansen, R.L. (2005, April). *Detecting CPI faking in a police sample: A cautionary note.* Paper presented at the 2005 Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles.

Holden, R.R., & Bookc, A.S. (2009). Using hybrid Rasch-latent class modeling to improve the detection of fakers on a personality inventory. *Personality and Individual differences, 47,* 185-190.

Holden, R. R., & Passey, J. (2010). Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and individual differences, 49,* 446-450.

Hopwood, C.J., Morey, L.C., Rogers, R., & Ewell, K. (2007). Malingering on the Personality Assessment Inventory: Identification of specific feigned disorders. *Journal of Personality Assessment, 88,* 43-48.

Hough, L.M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11,* 509-244.

Hough, L.M., Eaton, N.K., Dunnette, M.D., Kamp, J.D, McCloy, R.A. (1990). Criterion related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581-595.

Jackson, D.N. (1970). A sequential system for personality scale development. In C.D. Spielberger (Ed.), *Current topics in clinical and community psychology, Vol. 2* (pp. 61 – 96). New York: NY: Academic Press.

Jackson, D.N., & Messick, S. (1958). Content and style in personality-assessment. *Psychological Bulletin, 55,* 243-252.

Kaiser, H.F. (1970). A second-generation little jiffy. *Psychometrika, 35*, 401-415.

Katz, Y. J., & Francis, L. J. (1991). The dual nature of the EPQ Lie scale? A study among university students in Israel. *Social Behavior and Personality, 9,* 217–222.

Kuncel, N.R., & Bornemam, M.J. (2007). Toeard a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment, 15,* 220-231.

Kuncel, N.R., Borneman, M., & Kiger, T. (20012). Innovative Item Response Process and Bayesian Faking Detection Methods. In M. Ziegler, C. MacCann, & R. Roberts (Eds.) *New Perspectives on Faking in Personality Assessment* (pp.3-16) New York: Oxford University Press, Inc.

Lagerspetz, K.J.M., & Björkqvist, K. (1994). Indirect aggression in boys and girls. En L.R. Huesmann (Ed). *Aggressive behaviour: Current perspectives.(pp.131-150)* New York. Plenum.

Lajunen, T., & Scherler, H. R. (1999). Is the EPQ Lie scale bidimensional? Validation study of the structure of the EPQ Lie scale among Finnish and Turkish university students. *Personality and Individual Differences, 26,* 657–664.

Lautenschlager, G.J. (1986). Within-subject measures for the assessment of individual differences in faking. *Educational and Psychological Measurement, 46,* 309-316.

Leite, W. L., & Cooper, L. A. (2010). Detecting Social Desirability Bias Using Factor Mixture Models. *Multivariate Behavioral Research, 45*, 271- 293.

Loo, R. (1995). Cross-cultural validation of the dual nature of the EPQ Lie scale with a Japanese sample. *Personality and Individual Differences, 18,* 297–299.

McCrae, R.R., & Costa, P.T. (1983). Social Desirability Scales: More Substance tan Style. *Journal of Consulting and Clinical Psychology, 51*, 882-888.

McFarland, L.A., & Ryan, A.M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology,* 85, 812-821.

McFarland, L.A., & Ryan, A.M. (2006). Toward and Integrated Model of Applicant Faking Behavior. *Journal of Applied Social Psychology*, 36, 979-1016.

Mersmer-Magnus, J., & Viswevaran, C. (2006). Assessing response distortion in personality tests: A review of research designs

and analytic strategies. In R.L. Griffith & M.H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 85-113). Greenwich, CT: Information Age Publishing.

Michaelis, W., & Eysenck, H. J. (1971). The determination of personality inventory factor patterns and intercorrelations by changes in real-life motivation. *Journal of Genetic Psychology, 118,* 223–234.

Morales-Vives, F., Codorniu-Raga, M.J., & Vigil-Colet, A. (2005). Características psicométricas de las versiones reducidas del cuestionario de agresividad de Buss y Perry. *Psicothema, 17,* 96-100.

Muller-Hanson, R., Heggestad, E.D., & Thorsnton, G.C. (2003). Faking and selection: Considering the use of personality test from select-in and select-out perspectives. *Journal of Applied Psychology, 88,* 348-355.

Muñiz, J. (1979). Introducción a la teoría de respuesta a los ítems. Madrid: Piramide

Ones, D.S., Vieswesvaran, C., & Reiss, A.D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81,* 660-679.

Pannone, R.D. (1984). Predicting test performance: A content valid approach to screening applicants. *Personnel Psychology, 37,*507-514.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), Measures of personality and social psychological attitudes (pp. 17–59). San Diego: Academic Press.

Paulhus, D.L. (1998). *Paulhus Deception Scales: The Balanced Inventory of Social Desirable Responding-7 User's Manual.* North Tonawanda, NY: Multi-Healthy Systems Inc.

Paulhus, D.L. (2002). Socially desirable responding: The evolution of a construct. In H.I. Braum, D.N. Jackson, & D.E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ: Lawrence Erlbaum Associates.

Pauls, C.A., & Crost, N.W. (2005). Effects of different instructional sets on the construct validity of the NEO-PI-R. *Personality and individual differences, 39,* 297-308.

Rothstein, M.G., & Goffin, R.D. (2006). The use of personality measures in personnel selection: What does the current research support? *Human Resource Management Review, 16,* 155-180.

Schmit, M.J., & Ryan, A.M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78,* 966-974.

Schmitt, N., & Oswald, F.L. (2006). The impact of corrections for faking on the validity of noncognitive measures in selection settings. *Journal of Applied Psychology, 86,* 613-621.

Selby, M.J. (1984). Assessment of violence potential using measures of anger, hostility, and social desirability. *Journal of Personality Assessment, 48,* 531-544.

Smith, D.B., & Ellingson, J.E. (2002). Substance versus style: A new look at the social desirability in motivating context. *Journal of Applied Psychology, 87,* 211-219.

Suarez, E.C., & Williams, R.B. (1989). Situational determinants of cardiovascular and emotional reactivity in high and low hostile men. *Psychosomatic Medicine, 51,* 404-418.

Vaillancourt, T. (2005). Indirect aggression among humans: social construct or evolutionary adaptation?. En R. E. Tremblay., W.W. Hartup., y J.Archer (Eds.) *Developmental origins of aggression (pp.158-177)*. New York. Guilford Press.

Vigil-Colet, Ruiz-Pàmies, Anguiano-Carrasco, & Lorenzo-Seva. (2012). The impact of social desirability on psychometric measures of aggression. *Psicothema, 24,* 310-315.

Vigil-Colet, A., Morales-Vives, F., Lorenzo-Seva, U., Camps, E., & Tous, J. (in press). Development and validation of the Overall Personality Assessment Scale (OPERAS). *Psicothema*.

Viswesvaran, C., & Ones, D.S. (1999) Meta-analysis of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59,* 197-210.

Warren, P., Richardson, D.S., & McQuillin, S. (2011). Distinguishing Among Nondirect Forms of Aggression. *Aggressive Behavior, 37,* 1-11.

White, L.A., Nord, R.D., Mael, F.A., & Young, M.C. (1993). The assessment of Background and Life Experiences (ABLE). In T. Trend & J.H. Laurence (Eds.), *Adaptability screening for the armed forces* (pp. 101-162). Washington, D.C.: Office of Assistant Secretary of Defense (Force Management and Personnel).

White, L.A., Young, M.C., Hunter, A.E., & Rumsey, M.G. (2008). Lessons learned from transitioning personality measures from research to operational settings. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 291-295.

Zickar, M.J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied psychological measurement, 20,* 71-87.

Zickar, M.J., & Sliter, K.A. (2012). Searching for Unicorns: Item Response Theory-Based Solutions to Faking Problem. In In M. Ziegler, C. MacCann, & R. Roberts (Eds.) *New Perspectives on Faking in Personality Assessment* (pp.113-130) New York: Oxford University Press, Inc.

Ziegler, M., & Beuhner, M. (2009). Modeling social desirable responding and its effects. *Educational and Psychological Measurement, 69,* 548-565.

Ziegler, M., MacCann, C., & Roberts, R. (2012). Faking: Knowns, Unknowns and Points of Contention. In M. Ziegler, C. MacCann, & R. Roberts (Eds.) *New Perspectives on Faking in Personality Assessment* (pp.3-16) New York: Oxford University Press, Inc.

Ziegler, M., Toomela, A., & Beuhner, M. (2009). A reanalysis of Toomela (2003): Spurious measurement error as cause for common variance between personality factors. *Psychology Science Quarterly, 51,* 65-75.