A Multi-Sample Re-examination of the Factor Structure of Goldberg's IPIP 50-item Big Five Questionnaire

ABSTRACT

The factor structure of Goldberg's Big Five measures was examined via a confirmatory factor analytic (CFA) approach. Across seven samples, a CFA model, applied at the item level, in which two method bias factors indicating positive and negative item wording effects were estimated fit the data significantly better than a model without such item wording effects. Although orthogonal to the Big Five factors, the two item wording factors were positively correlated to each other across seven samples. Researchers using self-report measures to assess personality dimensions should consider applying models that include method bias factors.

Keywords:
Personality; Big Five structure; Confirmatory Factor Analysis

A Multi-Sample Re-examination of the Factor Structure of Goldberg's IPIP 50-item Big Five Questionnaire

Introduction

Personality, defined as, "individual characteristic patterns of thought, emotion, and behavior, together with the psychological mechanisms – hidden or not – behind those patterns" (Funder, 2001, p. 2) is commonly linked to work behavior and outcomes. Of the myriad ways of describing these complex patterns, the lexical approach has been used perhaps more than any other. This method assumes that personality attributes can be well-captured universally (i.e., using similar language across the world's many cultures). This notion is important for cross-cultural generalization in personality assessment.

The dominant model employed in most lexical studies within Northern European languages is a five factor structure, the Big Five, consisting of Extraversion, Agreeableness, Conscientiousness, Neuroticism (often measured as Emotional Stability), and Openness to Experiences (sometimes called Intellect). The Big Five has become the most well-known taxonomy of personality to date (Saucier & Goldberg, 2003). Correlations between summated scale scores on most Big Five personality tests are generally positive, leading several authors to suggest that the Big Five dimensions may not be orthogonal, but rather correlated indicators of higher order personality dimensions (e.g., Musek, 2007).

Some suggest that the Big Five dimensions are indicators of two higher order factors, with Agreeableness, Conscientiousness, and the inverse of Neuroticism as indicators of a Stability factor and Openness and Extraversion as indicators of a Plasticity factor (Digman, 1997; DeYoung, Peterson, & Higgins, 2001). Others suggest that there is one overriding personality factor, Evaluation (Goldberg & Somer, 2000; Saucier, 1997) or the Big One (Musek, 2007).

Although several personality tests have been developed around the Big Five conceptual model, most are only available at a cost (e.g., the NEO-PI, 16-PF, HPI, CPI) and thus are infrequently used by researchers and potentially too expensive for use by smaller organizations and researchers. The International Personality Item Pool (IPIP), developed by Lewis Goldberg, is an increasingly popular no-cost alternative to proprietary measures of these traditional five factors of personality. The 50-item version of the IPIP scales has been recently validated and shown to have good reliability and validity compared to established five factor measures of personality such as the NEO-FFI (Lim & Ployhart, 2006).

Despite the widespread use and acceptance of the type of five factor personality measures such as the IPIP scales, lingering and serious limitations in personality assessment continue to be highlighted by psychological researchers and practitioners. Perhaps the clearest recent illustration of this comes from the field of industrial-organizational psychology, where in spite of a resurgence in popularity of the use of personality tests in employment selection since the early 1990s, a recent review questions the appropriateness and utility of personality assessments for employment selection and other high-stakes testing situations (Morgeson, Campion, Dipboye, Murphy, & Schmitt, 2007). A major criticism of personality tests raised by these researchers and others is that these assessments rely predominantly on self-reported information. In such assessments, applicants are asked to endorse or rate their agreement with multiple statements of behavioral descriptions that supposedly underlie personality constructs (i.e., “I have little concern of others”; “I have a good imagination”). A potential consequence of relying on self-reported information is that the resulting scores may include variance due to the items and response format that cannot be explained by the *a priori* personality dimensions alone.

Common Method Bias

A major common concern in studies with self-report methodologies is the possibility of common method bias being responsible for substantive relationships when variables representing multiple dimensions are collected from the same source (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Specifically, the issue is that the observed covariances between variables of interest could be inflated or deflated by variance due to the method rather than to the underlying constructs or variables of interest.

The potential for measures of the Big Five traits to be influenced by common method bias was first reported by Schmit and Ryan (1993) who factor analyzed responses to individual items of the NEO-FFI (Costa & McCrae, 1989) within applicant and non-applicant samples in an organization. An exploratory factor analysis (EFA) of a non-applicant sample demonstrated the expected five-factor solution, but in the applicant sample, a six-factor solution fit the data best. Schmit and Ryan labeled this sixth factor an "ideal employee" factor, noting that it "included a conglomerate of item composites from across four of the five subscales of the NEO-FFI" (Schmitt & Ryan, 1993, p. 971). Interestingly, items from all five NEO-FFI subscales loaded on this factor, suggesting that the "ideal employee factor" represented a form of common method bias.

Additional studies (e.g., Frei, 1998; Frei, Griffith, Snell, McDaniel, & Douglas, 1997) comparing factor structures of the Big Five measures between faking good versus honest responding groups have also shown differences in the number of latent variables, error variances, and correlations among latent variables across groups. Recently, Biderman and Nguyen (2004) investigated a model in which a common method factor specifically representing the ability to distort or fake responses to personality items was included. In that application and subsequent

ones (Wrensen & Biderman, 2005; Clark & Biderman, 2006), individual differences in response distortion in faking responding groups were captured by a single latent variable similar to what Podsakoff and colleagues (2003) labeled an "unmeasured method" effect.

Apart from its potential nuisance effects on personality measurement, common method bias has also been found to be a substantive variable in the study of relationships between personality and work outcomes. For example, halo error as a common method bias has been found to relate to performance rating accuracy (Sulsky & Balzer, 1988).

Item Wording Effects

As discussed, the use of self-report questionnaires to measure personality is a common practice. Conventional wisdom suggests that it is necessary to include an equal number of negatively worded items (e.g., "I don't talk a lot") during scale development to reduce response bias such as acquiescence (Nunnally, 1978). In assessing Extraversion, for example, if a five-point response scale of agreement is used, then a "5" response on a positively worded item (e.g., I am the life of the party) should represent roughly the same amount of Extraversion as "1" for the negatively worded item (e.g., I don't talk a lot). Standard practice is to reverse-code responses to negatively worded items, so that large positive response value represent greater amounts of whatever construct is being measured regardless of item wording. This practice of using a variety of item wording formats, including negatively worded items to counteract respondents' acquiescence, can be found throughout most areas of organizational research including personality assessment (e.g., Paulhus, 1991; Motl & DiStefano, 2002; Quilty et al., 2006), leadership behavior (e.g., Schriesheim & Hill, 1981; Schriesheim & Eisenbach, 1995), role stress (Rizzo, House, & Lirtzman, 1970), job characteristics (Harvey, Billings, & Nilan, 1985), and organizational commitment (e.g., Meyer & Allen, 1984).

Researchers in personality assessment have long been aware of response bias due to acquiescence (Paulhus, 1991). Unfortunately, the negatively worded items that were introduced to counter response tendencies such as acquiescence have been found to be associated with systematic and construct irrelevant variance in scale scores. For example, Hensley and Roberts (1976) conducted an exploratory factor analysis (EFA) of the Rosenberg's Self-esteem scale and found the scale consisted of two factors: one loading on positively worded items and the other on negatively worded items. This finding was later replicated and the factors labeled positive and negative self-esteem (Carmine & Zeller, 1979). Later studies using CFA all showed that a model in which two method effects (one representing positively and one negatively worded items) were estimated provided the best fit to the data (e.g., Marsh, 1996; Tómas & Oliver, 1999).

Unfortunately, the inclusion of negatively worded items in leadership behavior measures has been shown to decrease a scale's reliability and validity (Schriesheim & Hill, 1981; Schriesheim & Eisenbach, 1995). In Schriesheim and Hill (1981), the authors examined the internal consistency estimates of the Leadership Behavior Description Questionnaire (LBDQ) – form XII Stogdill (1963), using all positively worded items versus negatively worded items versus a combination of both, to measure the leadership behavior of initiating structure. They found that the negatively worded items produced the lowest scale reliability, followed by a mix with all positively worded item scale having the highest reliability. Schriesheim and Eisenbach (1995) further found that a CFA model with one trait factor (i.e., initiating structure) and two item wording factors (positive and negative wording formats) provided the best fit to the data based on the chi-square difference test.

The role conflict and role ambiguity scale developed by Rizzo, House, & Lirtzman (1970) includes both positively and negatively worded items. A CFA model including a general

factor of role stress and a second orthogonal factor representing an item wording effect was found to provide the best fit to the data (McGee, Ferguson, & Seers, 1989). In another study using a Multitrait-Multimethod (MTMM) and variance partitioning approach, an item wording factor orthogonal to the substantive factors of role conflict and role ambiguity was found to explain 18% of the item variance in role conflict and 19% of the item variance in role ambiguity (Harris & Bladen, 1994).

An orthogonal item wording effect was also found to alter the factor structure of the Job Diagnostic Survey (JDS) developed by Hackman and Oldham (1975). In a study to replicate the factor structure of the JDS, Harvey and colleagues (1985) found that including a factor indicated by the negatively-worded items significantly increased the CFA model fit. They also found that negatively worded items contributed a substantial amount of construct irrelevant variance in this study (Harvey et al., 1985).

As one final example within organizational research, Magazine, Williams, and Williams (1996) found that negatively worded items complicated the interpretation of the factor structure of Meyer and Allen's (1984) organizational commitment scale. Specifically, the authors found that adding an orthogonal reverse coding factor representing the negatively worded item effect to the CFA model in addition to two substantive factors (i.e., affective commitment and continuance commitment) resulted in the best fit to the data. The factor loadings for the reverse-scored items were all significant while maintaining the significance of factor loadings to their respective substantive factors.

In sum, these existing organizational psychology studies have shown that adding one or two item wording factors orthogonal to the *a priori* substantive or trait factor(s) was often found to significantly increase the model fit. Further, negatively worded items were found to contribute

a substantial amount of variance irrelevant to the constructs of interest. Given the increasing usage of personality assessments in industrial and organizational psychology research and practice it is surprising that no attempts have been made to examine potential item wording effects on the factor structure of the IPIP scale.

Examination of the IPIP 50-item scale reveals 26 positively-worded and 24 negatively-worded items. Each subscale contains both positively- and negatively-worded items. The number of positively-worded items in the subscales is five, six, six, two, and seven for Extraversion, Agreeableness, Conscientiousness, Stability, and Openness respectively. Because of the prevalence of negatively-worded items, a purpose of the present study was to examine the need for separate method factors indicated by positively-worded items and negatively-worded items in modeling responses to the 50-item IPIP scale.

Goodness of fit

Investigation of item-wording factors requires the use of individual items as indicators of the factors. As mentioned above, Lim and Ployhart (2006) conducted the most extensive validation study of the IPIP to date, replicating the factor structure originally proposed by Goldberg (1999). However, their confirmatory factor analysis (CFA) model only achieved acceptable fit when parcels were used as indicators. Lim and Ployhart (2006) called for future research to replicate the factor structure of the IPIP at the individual item level. One possible reason for poor fit when individual items are used as indicators is that using a larger number of items increases the likelihood of model misspecification as those items may systematically share sources of common variance not specified *a priori*. This, in turn, may reduce the model fit (Little, Cunningham, Shahar, & Widaman, 2002). If, in fact, poor fit is due to unmodeled covariances between individual items, whether such misspecification affects the factor structure

of the personality measures is an empirical question. Thompson and Melancon (1996) reported no changes in the factor structure of the Personal Preferences Self-Description questionnaire as the number of items per parcel increased, although goodness-of-fit improved. McMahon and Harvey (2007) reported a substantial improvement in model fit of the Multidimensional Ethics Scale (MES) when modeled at the subscale/parcel level compared to when modeled at the item level. No comparable analyses have been performed on the IPIP Big Five scales.

The Present Study

While there has been much concern over the nuisance of common method bias, little research has been done on these issues as they pertain to personality assessment, especially in conditions in which participants were expected to respond honestly. To our knowledge, Roth and colleagues examined the effect of method bias on the relationships among several personality variables including conscientiousness, locus of control, and work ethic (Roth, Hearp, & Switzer, 1999). However, in that study, method bias was estimated in a series of CFAs in which the personality variables were modeled as parcels, rather than individual items. Only one preliminary study has investigated method bias and modeled it at the individual item level. In that study Biderman (2007) estimated a method bias factor in four datasets and found that models estimating a method factor exhibited better fit than models without a method factor. Other than the Biderman (2007) study, we are aware of no published studies examining whether common method variance affects the IPIP measure.

The present study addresses three important gaps in the literature. First, we wanted to see if common method variance exists in the widely used IPIP measure of the five factor model of personality. As mentioned above, this represents an extension of Biderman's (2007) study. Given the substantial evidence of the importance of method bias in a variety of studies involving self-

report questionnaires, we expect that it also plays a role in responses to Big Five questionnaires. Thus,

Hypothesis 1: Estimating the method effect in addition to the five *a priori* constructs will significantly improve the CFA model fit when modeled at the individual item level.

Second, given the presence of method bias, we wanted to examine whether there are also item-wording effects. Because of the large number of studies reporting differences in bias involving positively-worded items vs. negatively-worded items, we also expected an improvement in goodness of fit when estimating two method biases as opposed to one. Thus,

Hypothesis 2: Estimating the item wording method effect(s) in addition to the five *a priori* constructs underlying the IPIP data will significantly improve the CFA model fit when modeled at the individual item level.

Third, we wanted to demonstrate the consistency of these effects across multiple samples (seven, to be exact). By addressing these objectives, the present study also extends Lim and Ployhart's (2006) validation effort in increasing model fit by modeling individual items, rather than parcels. If evidence for the existence of item-wording method factors is found, any future models of the IPIP subscales will need to use items as indicators when taking item wording effects into account. The goodness-of-fit of the models presented here will serve as an indicator of what future investigators could expect.

Method[1]

Participants

[1] In interests of full disclosure, we note that some of the datasets reported upon here and some of the analyses on these single datasets have been reported in other venues mentioned in the sample description.

In the present study, we report the results of a CFA model with and without method effects shown in Figures 1 through 3 using data from seven separate samples described in detail below.

Sample 1: 203 undergraduate and graduate business students at a Mid-Atlantic University – United States participated in exchange for partial course credit in spring 2001. The sample was 86 male and 117 female, with a mean age of 25.33 years (*SD* = 6.24). By ethnicity, the sample was fairly diverse with 55.7% White, 24.1% Black, 14.8% Asian, 1% Hispanic, and 4.4% reporting "other". Other aspects of these data were reported in Nguyen, Biderman, & McDaniel (2005), Biderman & Nguyen (2004), and Biderman (2007).

Sample 2: 166 undergraduate students enrolled in an introductory psychology course at a southeastern university in the United States. The sample was 55 males, with a mean age of 23.4 (*SD* = 7.8) and 110 females with mean age of 21.7 (SD = 5.5). There were 58.9% White, 29.4% African American, 4.9% Hispanic and 6.8% "other" (Wrensen & Biderman, 2005; Biderman, 2007).

Sample 3: 360 students undergraduates with158 males with mean age 22.4 (SD=8.5) and 202 females with mean age 23.6 (SD = 12.3). Ethnicity was 77.7% White, 19.4% African American, and 2.9% "other" (Damron, 2004; Biderman, 2007).

Sample 4: 185 undergraduate students enrolled in an introductory psychology course at a southeastern university in the United States. The sample was 71 male, with an average age of 19.39 years (*SD* = 2.65). By ethnicity, 59.5% were White, 33% were Black, 3.2% were Asian or Pacific Islander, 2.2% were Hispanic, and 2.1% were Native American and/or other (Biderman, Nguyen, & Sebren, 2007; Biderman, 2007).

Sample 5: Participants were 764 employees of a national private personal finance company with job titles of "Sales Associate" or "Sales Manager". Eighty-six percent were female; 59% were White, 24% Black, 9% Hispanic and 8% described themselves as "Other". The essential duties of each job were the same with respect to interacting with customers. Each job required the incumbent to perform duties and tasks in the areas of selling, customer service, and debt collections. Participants were asked to complete the IPIP-50 item version presented using a web-based computer system (Biderman, Nguyen, Mullins, & Luna, 2008).

Sample 6: Participants were 311 undergraduate students from seven separate classes (six at a large Midwestern university and one at a medium-sized university in the eastern United States). The IPIP data were collected as part of a larger study of work-related stress and performance. Of these 311 participants, 35.7% were male. The average age was about 21 years. All participation was voluntary; though completion of both surveys earned participants a small amount of course credit and an entry into a raffle for one of several Amazon.com gift certificates (Cunningham, 2007)

Sample 7: Participants were 404 undergraduates enrolled at the University of Tehran. The responding of all students to the procedures of this project was voluntary, completely anonymous, and in conformity with institutional ethical guidelines. Questionnaires were administered in classroom settings to groups of varying sizes. Mean age of all participants was 21.5, and 63.4% were female.

Procedure

The personality measure used in all seven samples was the 50-item version from the IPIP (Goldberg, 1999). For Sample 7, items were translated into Persian, then back-translated into English by an individual not previously involved in the translation procedures. Noteworthy

discrepancies between the original and back-translated English statements were rare and successfully resolved through appropriate revision of the Persian translation.

In all samples, participants were instructed to respond honestly to the IPIP-50 item version. Participants were asked to endorse items reflecting what they thought of themselves at the time, not how they wished to be in the future. Anchors of items ranged from "1" = very inaccurate to "5" = very accurate. For dataset 4, the response scale ranged from "1" = very inaccurate to "7" = very accurate. Reliability estimates of summated scales for the five dimensions are shown in Table 3.

Analyses

All CFA models were estimated using Mplus V4.2 (Muthén & Muthén, 1998-2006). Model 1 contained five latent variables representing the *a priori* Big Five constructs of extraversion, agreeableness, conscientiousness, emotional stability, and openness/intellect respectively. Each item loaded on the appropriate latent variable. Correlations among the latent variables were estimated. Thus, Model 1 was a standard CFA model of the IPIP 50-item version with items as indicators of the latent variables (See Figure 1).

Model 2 was identical to the first with the exception that a sixth latent variable, labeled M, was included. All 50 items were required to load on M. For purposes of model identification M was constrained so that it was orthogonal to all of the Big Five factors (Williams, Ford, & Nguyen, 2002). Thus, the Method factor, M, represented variance shared among all 50 items over and above any variation attributable to the *a priori* Big Five constructs. Model 2 is analogous to that presented in cell 3A in Table 4 of Podsakoff et al. (2003) where the latent variable is called an "unmeasured latent methods factor" (See Figure 2).

Model 3 was identical to the second model, except that the Method factor was split into two factors: one indicated by positively worded IPIP items (Mp) and one indicated by negatively worded IPIP items (Mn) (See Figure 3).

Insert Figures 1, 2 & 3 about here

We used various goodness-of-fit statistics for model evaluation. We reported the Chi-square statistic, Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA); and the Standardized Root Mean Square Residual (SRMR). As noted in prior research, whereas RMSEA was found to be most sensitive to misspecified factor loadings (a measurement model misspecification); SRMR was found to be most sensitive to misspecified factor covariances (a structural model misspecification) (Hu & Bentler, 1999). Later studies replicating Hu and Bentler's seminal work confirmed that SRMR and RMSEA values were found to perform better than other fit indexes at both retaining a correctly specified (i.e., true) model and rejecting a misspecified model (Sivo, Fan, Witta, & Willse, 2006). Thus, both values are reported in this study. Whereas models with CFI values close to .95 are having a good fit to the data, RMSEA values less than .06 and SRMR values less than .08 are considered acceptable fit (Hu & Bentler, 1999).

Results

Table 1 presents the above-mentioned fit statistics of three models applied to seven datasets. Hypothesis 1 was that estimating the method effect in addition to the five *a priori* constructs would significantly improve the CFA model fit when modeled at the individual item level. Because all three models were nested, differences in model fit were tested using chi-square difference tests. As shown in Table 1, across all seven samples Model 2 (in which a common

method factor was estimated) had a significantly better fit to the data than Model 1 (no method factor), $\Delta\chi^2(50)$ ranges from 205.24 to 683.55, all significant at $p < .001$. The CFIs from Model 1 were lower (ranging from .62 to .78 with a mean of .70) than for Model 2 (ranging from .69 to .83 with a mean of .76) across the seven samples.

Both the RMSEA and SRMR also consistently indicated better fit for Model 2 than Model 1 (ranging from .05 to .08 with a mean of .07 for Model 1 and .05 to .07 with a mean of .06 for Model 2). The SRMR values ranged from .07 to .10 with a mean of .09 for Model 1 and .05 to .08 with a mean of .07 for Model 2 respectively. Taken together, these fit indices indicated that common method bias was needed to explain the IPIP data. Thus, Hypothesis 1 was fully supported.

Hypothesis 2 stated that estimating the item wording method effect(s) in addition to the five *a priori* constructs underlying the IPIP would significantly improve the CFA model fit when modeled at the individual item level. The chi-square difference test revealed that Model 3 in which two method factors were estimated (one indicated by positively worded items and one indicated by negatively worded items) had a better fit than Model 2 across the seven samples, $\Delta\chi^2(1)$ range from 22.91 to 346.58, $p < .001$.

In terms of fit indices, Model 3 had a higher CFI (ranging from .71 to .85 with a mean of .78) than did Model 2 (ranging from .69 to .83 with a mean of .76) across the seven samples. Both the RMSEA and SRMR showed Model 3 fit the data better than Model 2 across 7 samples although the mean values of these fit statistics changed only in the second decimal place. Specifically, the RMSEA values ranged from .043 to .070 with a mean of .058 for Model 3 and .045 to .074 with a mean of .062 for Model 2 respectively. The SRMR values ranged from .043 to .088 with a mean of .069 for Model 3 and .047 to .084 with a mean of .071 for Model 2

respectively. These fit indices indicated that the effect of item wording format needed to be accounted for in modeling IPIP data adequately. Thus, Hypothesis 2 was supported.

Although Mp and Mn were estimated orthogonal to the Big Five latent variables, they were allowed to correlate with each other. Those correlations for the seven datasets were .77, .84, .75, .75, .81, .33, and .45 respectively. All were significantly different from 0, $p < .001$ for all.

Insert Tables 1, 2, 3, & 4 about here

Table 2 shows the observed and latent factor correlations of the IPIP scales as applied in three CFA models to the seven samples. Table 3 shows the reliability estimates of observed and latent variables as modeled in the three CFAs applied to the seven datasets. As shown in Table 2, across seven samples, the mean intercorrelations among the Big Five observed scale scores ranged from .08 (between Extroversion and Conscientiousness) to .30 (between Extroversion and Openness to Experience/Intellect) with a grand mean of .21. This mean value is consistent with, *albeit* a bit higher than what Lim and Ployhart's (2006) reported (r =.16) in their previous IPIP scale validation study.

A further examination of Table 2 reveals that the intercorrelations of the Big Five latent variables (i.e., factor correlations) were higher than their observed scale counterparts although when method effects were added to the model, these relationships either decreased or became negative. For example, in Model 1 where no method effect was estimated in the model, across seven samples, the mean factor correlations of the Big Five traits ranged from .12 (between Extroversion and Conscientiousness) to .40 (between Extroversion and Openness/Intellect) with a grand mean of .27. In Model 2 where a common method factor was estimated, the mean factor

correlations of the Big Five were reduced to between -.15 (between Agreeableness and Conscientiousness) and .28 (between Agreeableness and Conscientiousness) with a grand mean of .07. It should also be noted that two mean factor correlations (between Extroversion and Emotional Stability and between Agreeableness and Emotional Stability) actually became negative when a method factor was estimated.

A similar pattern of results was found with Model 3. Specifically, when two method factors were estimated to account for positive and negative item wording effects in the IPIP scales, the mean factor correlations of the Big Five ranged from -.35 (between Agreeableness and Emotional Stability) to .24 (between Extroversion and Openness) with a grand mean of -.01. Again, it is noted that four factor correlations became negative when two method factors were estimated (See Table 2).

As shown in Table 3, the internal consistency reliabilities of the IPIP scales were the lowest when estimated as observed variables. Specifically, Cronbach's alpha estimates ranged from .74 to .91 with a mean of .85 for Extroversion; .67 to .84 with a mean of .79 for Agreeableness; .71 to .85 with a mean of .80 for Conscientiousness; .80 to .89 with a mean of .85 for Emotional Stability; .69 to .81 with a mean of .76 for Openness/Intellect. These alpha coefficients are consistent with those reported by Goldberg via the official IPIP site (http://ipip.ori.org/newBigFive5broadTable.htm) and by Lim and Ployhart (2006). When estimated as latent variables, all reliability estimates were higher across the three CFA models. Specifically, across the seven samples, the mean reliability estimate of four out of five Big Five constructs showed a substantial increase (Extroversion from .85 to .9 ranges; Agreeableness from .79 to .9 ranges, Conscientiousness from .8 to .9 ranges, and Openness from .76 to .9 ranges). Only Emotional Stability did not show a consistent pattern of increase in reliability, ranging from

.85 when estimated as an observed variable to .89 when estimated as a latent variable in Model 1 (no method effect model), but decreasing to .78 when estimated as a latent variable in Model 2 (method effect model) and then increasing to .83 when estimated as a latent variable in Model 3 (item wording effect model).

Table 4 shows the amount of variance explained by the Big Five or substantive dimensions, method, and random error respectively by competing CFA models applied to seven datasets. We followed Williams, Cote, and Buckley's (1989) procedure in partitioning variances explained by each set of factors using standardized factor loadings. To be consistent with the Multi-trait-Multi-method (MTMM) literature, in this study, the term "trait" was used interchangeably with "substantive". As shown in Table 4, the amount of variance explained by the Big Five traits decreased from Model 1 (no method Model) to Model 2 (one Method factor Model), and Model 3 (two Method factor Model). Specifically, for Model 1 where method effects were assumed to be zero, across seven samples, trait variance ranged from 24.2% to 39.3% with a mean of 33.3%. However, for Model 2 where one method effect was estimated, trait variance decreased to ranging from 16.5% to 38% with a mean of 26.4% across seven samples. Method variance ranged from 6% to 14.1% with a mean of 10.1% across seven datasets.

The amount of variance explained by the Big Five traits was reduced to the lowest in Model 3 where two method factors were estimated, one for positively and one for negatively worded factors. Specifically, trait variance ranged from 16.2% to 31% with a mean of 23.8% across seven samples. Method variance, in contrast, increased from Model 2 to Model 3 with a range from 7.3% to 17.7% and a mean of 14.1%. The amount of variance explained by random error was fairly high even after partialling out trait and method variance. Across seven samples,

error variance ranged from 51.3% to 73.7% with a mean of 62.1%. This finding was consistent with previous research in psychological assessment (Harris & Bladen, 1994).

Discussion

The purpose of this study was to investigate whether common method variance exists in IPIP data and to investigate whether modeling method effects specific to item wording format of the IPIP scales explained the data more adequately than models of method effects that ignored item wording. Overall, Model 3 where common method variance in the form of item wording effects was estimated was considered the best fitted model to the IPIP data across seven samples based on fit statistics, reliability estimates, and factor loadings. Although method factors explained less than 20% of the variance in the IPIP items, this amount was enough to inflate the percent of variance attributed to the Big Five traits (i.e., factor correlations) in Model 1, where method variance was assumed to be zero. That is, excluding method effects from Model 1 resulted in a misspecification, causing the variance that would normally have been due to method to be, instead captured by correlations among the Big Five dimensions. When method factors were introduced in Models 2 and 3, the percentage of variance attributed to traits was reduced to its true value. This finding was consistent with previous research on method variance being responsible for inflating substantive relationships (e.g., Doty & Glick, 1998).

Further, we found that item wording format should be taken into account when modeling IPIP scales at the item level. It is important to note that although both the RMSEA and SRMR values for Model 3 in our study met or exceeded the recommended cutoff (e.g., Hu & Bentler, 1999); the CFI values (ranging from .70 to .85) were less than desired based on the traditional cutoff of .95 (e.g., Hu & Bentler, 1999). We note that 23% of the variation in CFI values was explained by variation in sample sizes such that larger CFI values tend to accord larger sample

sizes – other things equal (Sivo et al., 2006). In our study, the largest CFI value (.85) in Model 3 was that of Sample 5 with more than 700 cases. Thus, the lack of fit indicated by lower than desired CFI values should be considered in connection with the indications of fit provided by our reported RMSEA and SRMR values (e.g., Brown, 2006).

Possible Reasons for Lack of Fit

We demonstrated with data from seven separate samples that model fit for the IPIP measure of the Big Five personality traits could be improved substantially with the addition of item wording effects. Even with this improvement, however, the fit was only considered acceptable based on SRMR and RMSEA values. The CFI still has room for improvement based on conventional cutoff of .95 recommended in previous studies (e.g., Hu & Bentler, 1999). We offer two potential reasons for this continued lack of fit indicated by CFI. The first pertains to the way negatively worded items are phrased. For example, within the IPIP there are two types of negatively worded item formats: polar opposite (e.g., "I am easily disturbed") and negated regular (e.g., "I don't talk a lot"). Just as a single method factor did not represent the positively and negatively worded items as well as separate factors for each wording, it may be that Mn did not represent these two types of negatively worded items as well as separate Mn factors would have. Several studies in leadership behavior have shown that polar opposite and negated polar opposite wording were found to cause the most harmful effect on scale reliability and validity (e.g., Schriesheim & Eisenbach, 1995; Schriesheim et al., 1991) because it is difficult to create negatively worded items to reflect the same meaning of the positively worded counterparts (Rorer, 1965).

The second possible explanation for the continued lack of model fit, even after including our hypothesized method factors, is the possible carelessness of respondents or lack of self-

insight. One study reported that careless responding by only 10% of respondents could be enough to result in a construct-irrelevant factor from a CFA using non-regularly worded items (Schmitt & Stults, 1985). Even if this is the case, however, a lingering problem is how to identify either careless responders, non-regularly worded items, or both. One way of identifying careless responders might be to use consistency of responding to items within a dimension as a measure. For example, Biderman (2007) investigated the use of scale standard deviations as indicators of consistency of responding. Non-regularly worded items, on the other hand, might be identified through consistently small loadings on the Big Five dimensions not accompanied by equally small loadings on method factors across studies. These would indicate items that were not influenced by the Big Five traits but that were subject to method biases, effects which probably do not depend on specific wording as much as do dimension influences.

We note that the factor correlations of the Big Five either decreased to near zero or became negative in Model 3. Such correlations have implications for conceptualizations of the Big Five that posit higher order factors indicated by the Big Five factors. For example, a model proposed by Digman (1997) and DeYoung (2001) suggest that Agreeableness, Conscientiousness, and Emotional Stability together indicate a higher order factor called Stability. In our Model 3, however, the mean correlation of Agreeableness and Conscientiousness across the seven datasets was -.08; that of Agreeableness and Emotional Stability was -.35 and that of Conscientiousness and Stability was -.02. These results do not support the Stability factor conceptualization.

The other higher order factor proposed by Digman (1997) and DeYoung et al (2001), Plasticity, is assumed to be indicated by Extraversion and Openness. The mean correlation between these two factors from our Model 3 was .24. This does provide some support for a

possible higher-order factor influencing these two personality dimensions. Certainly, our observed patterns of correlations provide little evidence for a single higher order factor (Musek, 2007); especially after item wording effect is taken into account. Of the 10 possible correlations between the Big Five dimensions, only four were positive while six were either negative or zero to two decimal places. This finding coupled with the fact that the amount of trait variance was the smallest in Model 3 as discussed earlier further confirmed that the Big Five traits were fairly independent constructs. That they are positively correlated as reported in prior research (e.g., Musek, 2007) may be an artifact because method variance was not estimated.

We point out that the models presented here are quite different conceptualizations from those assuming substantive higher order factors. Ultimately both types of models are attempts to account for item covariances from different Big Five dimensions. The higher order factors influence item responses only through the first order Big Five dimensions, and models assuming higher order factors account for item covariances from different dimensions by assuming that the Big Five dimensions themselves are correlated. Thus the accounting is through the Big Five dimensions. On the other hand the method effects proposed in the present models influence item responses directly and account for covariances between items across dimensions directly through the loadings on the method bias factors, bypassing the Big Five factors.

Because higher order factor models can fit the data no better than the model assuming freely estimated correlations between the lower order factors, i.e., Model 1 in the present study, the differences in goodness of fit reported above clearly favor the method bias models presented here. Their fit was better than the fit of Model 1 whose fit would be as good as or better than the fit of any model with one or higher order factors. Although our initial inclination based on the goodness-of-fit results is to reject the models assuming higher order factors in favor of a

different interpretation involving method effects, we note that it is possible that the correlations between the Big Five latent variables were negatively biased due to biases in the maximum likelihood estimation versus other estimation methods such as GLS - Generalized Least Square (Fan & Sivo, 2005). Thus any conclusions regarding correlations between the Big Five latent variables and rejection of consideration of higher order factors based on the estimates of those correlations reported here should be treated as tentative.

Because the negatively worded items were reverse-scored for all the studies reported here, the two factors, Mp and Mn, are defined so that a person who biases his/her responses so as to present himself/herself in a positive light on negatively worded items will have a high positive value on Mn. Thus high values of both Mp and of Mn represent distortions of self reported positions on the Big Five dimensions consistent with creation of a favorable impression. The high correlations between Mp and Mn found for five of the seven studies suggest that it may be possible to ignore the differences between Mp and Mn and estimate just one method bias latent variable. Or it might be desirable to treat Mp and Mn as indicators of a higher order method bias. Either procedure would produce a single method factor whose substantive value might be of interest. As Morgeson et al (2007) suggested; faking as a method factor may help to explain relevant criterion variance, noting ". . . self-monitoring is probably a good thing in most social context, suggesting that whatever contributes to faking may also contribute to job performance – especially when one employs a supervisory rating as the criterion as is so often the case" (p. 708). Indeed, in the previous analyses of the data of Sample 5, it was found that when M was included in a model along with the Big Five latent variables, it was the best predictor of supervisor ratings on three different performance dimensions (Biderman et al., 2008).

Our two method factors, Mp and Mn, may also be of substantive interest treated separately. The positive correlations between them suggest that persons tend to self-present in consistently across all items. It should be noted, however, that these correlations were not perfect, indicating that there are some differences in the tendency between the two types of item wording, especially for Samples 6 and 7.

Identifying situations that moderate the correlation between the two tendencies appears to be an interesting future research question. Moreover, identifying variables which correlate with one but the not other is also an area of interest. For example, Quilty, Oakman, & Risko (2006) found that a method factor indicated by negatively-worded items from the Rosenberg Self Esteem scale correlated positively with both Conscientiousness and Emotional Stability scales from both the 50-item and 100-item version of the IPIP measure while correlations with Mp from the self-esteem measure were negligible.

There also clear implications of the present results reported for use of scale scores to represent the Big Five dimensions. Specifically, these results suggest that an observed scale score will be a mixture of the characteristics of the Big Five dimension the score is supposed to represent and the test-taker's bias in responding to the items of the scale. If the scale is made up of primarily negatively-worded items, the scale score will be contaminated mostly with Mn. If it is a scale made up of primarily positively-worded items, the scale will be contaminated mostly with Mp. At the best, the consequences of such contamination will result in observed scores that are "noisier" than would be desired. Such noise may suppress correlations between the contaminated variables and other variables. For example, in a previous analysis involving Sample 4 in the present study, the correlation between Conscientiousness and an objective measure of academic performance went from .09 ($p > .05$) when the measure of

Conscientiousness was contaminated by M to .20 ($p < .05$) when an uncontaminated Conscientiousness measure was considered (Biderman et al., 2007).

We see two options for those desiring to improve the measurement of the Big Five traits by leveraging our approach to removing the contamination due to method bias from the IPIP item scores. The first would be to apply a measurement model estimating M or Mp and Mn and then adding whatever structural model representing the research question to that measurement model forming a structural equation model. The second would be to apply a measurement model estimating M or Mp and Mn and then compute factor scores of those Big Five (or method factor) dimensions representing the research question and use those factor scores to investigate the research question using common regression techniques. Note that both of these strategies would involve administration of most of the personality attributes be it the Big Five or others such as locus of control or work ethic even though only one personality dimension might be of interest because M and Mp and Mn are only estimable from a multi-dimensional model. Given the pervasiveness of M and Mp and Mn in the seven datasets reported upon here, it is difficult for us to envision situations in which summated scores are not contaminated by such effects.

Conclusions

CFA models in which method bias or biases were estimated were applied to the data of seven studies in which participants had responded to the 50-item IPIP questionnaire. In all datasets, a model containing a single method bias factor was found to fit the data significantly better than a model without a method factor. Moreover, for all datasets, a model with two method bias factors – one indicated by positively worded items and one indicated by negatively-worded items fit data the best. These results suggest that researchers using self-report

questionnaires to assess personality dimensions should seriously consider applying models that include method bias factors.

Method bias has been an aspect of responses to personality and other questionnaires of which investigators have been long aware but at the same time has been long neglected. The results of this study probably apply to other measures of the Big 5 as well pending future research. Perhaps it is now time to bring method bias out of the category of nuisance variable and examine its potential to provide information about personality that is not available in the summated scale-based measures that have been the purview of psychologists for nearly a half century.

REFERENCES

Biderman, M. D. 2007. *Method variance and Big Five correlations*. Paper presented at the 7th annual conference of the Association for Research in Personality. Memphis, TN.

Biderman, M. D., & Nguyen, N. T. 2004. *Structural equation models of faking ability in repeated measures designs.* Paper presented at the 19th Annual Society for Industrial and Organizational Psychology Conference, Chicago, IL.

Biderman, M. D., & Nguyen, N. T. 2006. *Measuring response distortion using structural equation models.* Paper presented at the conference, New Directions in Psychological Measurement with Model-Based Approaches. Georgia Institute of Technology, Atlanta, GA. February.

Biderman, M. D., Sebren, J., & Nguyen, N. T. 2007. *Time on task mediates the conscientiousness-performance relationship.* Paper presented at the 22nd Annual Conference of The Society for Industrial and Organizational Psychology, New York, NY. April.

Biderman, M. D., Nguyen, N. T., Mullins, B., & Luna, J. 2008. *A method factor predictor of performance ratings.* Paper accepted for presentation at the 23rd annual conference of The Society for Industrial and Organizational Psychology, San Francisco, CA.

Brown, T. A. 2006. *Confirmatory factor analysis for applied research.* New York: The Guilford Press.

Carmine, E. G., & Zeller, R. A. 1979. *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Clark III, J. M., & Biderman, M. D. 2006. *A structural equation model measuring faking propensity and faking ability.* Paper presented at the 21st annual conference of the Society for Industrial and Organizational Psychology. Dallas, TX - May.

Costa, P.T., & McCrae, R.R. 1989. *The NEO PI/FFI manual supplement*. Odessa, FL: Psychological Assessment Resources.

Cunningham**,** C. J. L. 2007. Need for recovery and ineffective self-management. *Dissertation Abstracts International:* Section B: The Sciences and Engineering, 68(4-B), 2695.

Damron, J. 2004. *An examination of the fakability of personality questionnaires: Faking for specific jobs.* Unpublished master's thesis. University of Tennessee at Chattanooga. Chattanooga, TN.

DeYong, C. G., Peterson, J. B., & Higgins, D. M. 2001. Higher-order factors of the big five predict conformity: Are there neuroses of health? *Personality and Individual Differences, 33*: 533-552.

Digman, J. M. 1997. Higher order factors of the Big Five. *Journal of Personality and Social Psychology, 73*: 1246-1256.

Doty, D. H., & Glick, W. H. 1998. Common methods bias: Does common methods variance really bias results? *Organizational Research Methods, 1,* 374–406.

Fan, X., & Sivo, S. A. 2005. Sensitivity of Fit Indexes to Misspecified Structural or Measurement Model Components: Rationale of Two-Index Strategy Revisited. *Structural Equation Modeling***,** *12*: 343-367.

Frei, R.L. 1998. *Fake this test! Do you have the ability to raise your score on a service orientation inventory*. University of Akron. Unpublished doctoral dissertation.

Frei, R.L., Griffith, R.L., Snell, A.F., McDaniel, M.A., & Douglas, E.F. 1997. *Faking of non-cognitive measures: Factor invariance using multiple groups LISREL*. Paper presented at the 12th Annual Meeting of the Society for Industrial & Organizational Psychology: St. Louis, MO.

Funder, D.C. 2001. *The personality puzzle* (2nd ed.). New York: Norton.

Goldberg, L. R. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe*, Vol. 7 (pp. 1-28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R., & Sommer, O. 2000. The hierarchical structure of common Turkish person-descriptive adjectives. *European Journal of Personality, 14*: 497-531.

Hackman, J. R., & Oldham, G. R. 1975. Development of the Job Diagnostic Survey. *Journal of Applied Psychology, 60*: 159-170.

Harris**,** M. M. & Bladen, A. 1994. Wording effects in the measurement of role conflict and role ambiguity: A multitrait-multimethod analysis. *Journal of Management, 20:* 887-901.

Harvey, R. J., Billings, R. S., & Nilan, K. J. 1985. Confirmatory factor analysis of the Job Diagnostic Survey: Good news and bad news. *Journal of Applied Psychology, 70*: 461-468.

Hensley, W. E., & Roberts, M. K. 1976. Dimensions of Rosenberg's Self-esteem scale. *Psychological Reports, 78*: 1071-1074.

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

Lim, B-C., & Ployhart, R. E. 2006. Assessing the Convergent and Discriminant Validity of Goldberg's International Personality Item Pool: A Multitrait-Multimethod Examination. *Organizational Research Methods, 9*, 29-54.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. 2002. To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, *9*, 151–173.

Magazine, S.L., Williams, L.J., & Williams, W.L. 1996. A confirmatory factor analysis examination of reverse coding effects in Meyer and Allen's affective and continuance commitment scales. *Educational and Psychological Measurement, 56*, 241-250.

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

McGee, G.W., Ferguson, C.E.Jr., & Seers, A. 1989. Role conflict and role ambiguity: Do the scales measure these two constructs? *Journal of Applied Psychology, 74*, 815-818.

McMahon, J.M.; & Harvey, R.J. 2007. The Psychometric properties of the Reidenbach-Robin Multidimensional Ethics scale. *Journal of Business Ethics, 72:* 27-39.

Meyer, J., & Allen, N. 1984. Testing the "Side-bet theory" of organizational commitment: Some methodological considerations. *Journal of Applied Psychology, 69*: 372-378.

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

Motl, R. W., & DeStefano, C. 2002. Longitudinal invariance of self-esteem and method effects associated with negatively worded items. *Structural Equation Modeling, 9*, 562-578.

Musek, J. 2007. A general factor of personality: Evidence for the Big One in the five-factor model. *Journal of Research in Personality, 41*: 1213-1233.

Muthén, L.K., & Muthén, B.O. 1998-2006. Mplus User's Guide. Fourth Edition. Los Angeles, CA: Muthén & Muthén.

Nguyen, N. T., Biderman, M. D., & McDaniel, M. 2005. Effects of response instructions on faking a situation judgment test. *International Journal of Selection and Assessment, 13*, 250-260.

Nunnally, J.C. 1978. *Psychometric theory*, 2nd ed. New York: McGraw-Hill.

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

Podsakoff, P.M., MacKenzie, S. B., Lee, J., & Podsakoff, N.P. 2003. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879-903.

Quilty, L.C.; Oakman, J.M.; & Risko, E. 2006. Correlates of the Rosenberg Self-Esteem Scale method effects. *Structural Equation Modeling, 13,* 99-117.

Rizzo, J. R., House, R. J., & Lirtzman, S. I. 1970. Role conflict and ambiguity in complex organizations. *Administrative Science Quarterly, 15*: 150-163.

Rorer, L.G. 1965. The great response style myth. *Psychological Bulletin, 63*: 129-156.

Roth, P. L., Hearp, C., & Switzer, F. S. III. 1999. The effect of method variance on relationships between the work ethic and individual difference variables. *Journal of Business and Psychology, 14:* 173-186.

Saucier, G. 1997. Effects of variable selection on the factor structure of person descriptors. *Journal of Personality and Social Psychology, 73*: 1296-1312.

Saucier, G., & Goldberg, L.R. 2001. Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of Personality, 69*, 847-879.

Saucier, G., & Goldber, L.R. 2003. The Structure of Personality attributes. In M.R. Barrick and A.M. Ryan (Eds.). *Personality and Work* (1st Ed.). Jossey-Bass: San Francisco, CA.

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., & Stults, D.M. 1985. Factors defined by negatively worded items. The results of careless respondents? *Applied Psychological Assessment, 9*, 367-373.

Schriesheim, C.A., & Hill, K.D. 1981. Controlling acquiescence response bias by item reversals: The effect of questionnaire validity. *Educational and Psychological Measurement, 41,* 1101-1114.

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

Schriesheim, C.A., & Eisenbach, R.J. 1995. An exploratory and confirmatory factor analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *Journal of Management, 21*, 1177-1193.

Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. 2006. The search for "optional" cutoff properties: Fit index criteria in structural equation modeling. *Journal of Experimental Education, 74*: 267-288.

Stogdill, R. M. 1963. *Manual for the leader behavior description questionnaire – Form XII*. Columbus: Bureau of Business Research, Ohio State University.

Sulsky, L.M., & Balzer, W.K. 1988. Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*, 497-506.

Thompson, B., & Melancon, J. G. 1996. Using item 'testlets' / 'parcels' in confirmatory factor analysis: An example using the PPSDQ-78. Paper presented at the annual meeting of the Mid-South Educational Research Association, Tuscaloosa, AL: November.

Tomás, J. M., & Oliver, A. 1999. Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling, 6*, 84-98.

Tull, K.T. 1998. *The effects of faking behavior on the prediction of sales performance using the Guilford Zimmerman Temperament Survey and the NEO Five Factor Inventory.* Unpublished Doctoral Dissertation. University of Akron.

Williams, L. J.; Cote, J.A., & Buckley, M.R. 1989. Lack of method variance in self-reported affect and perceptions at work: Reality or artifact? *Journal of Applied Psychology, 74:* 462-468.

Williams, L. J., Ford, L. R., & Nguyen, N.T. 2002. Basic and Advanced Measurement Models for Confirmatory Factor Analysis. In S. Rogelberg (Ed.). *Handbook of Research Methods in Industrial and Organizational Psychology* (pp.366-389). Oxford: Blackwell.

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

Table 1. Fit statistics of alternative CFA models applied to 7 datasets

	χ^2			*df*			*p*	CFI			RMSEA			SRMR		
	M1	M2	M3	M1	M2	M3		M1	M2	M3	M1	M2	M3	M1	M2	M3
Sample 1	2252.12	2031.74	1972.3	1165	1115	1114	.00	.73	.77	.79	.068	.064	.062	.09	.074	.072
Sample 2	2315.73	2048.11	2025.2	1165	1115	1114	.00	.64	.70	.71	.077	.071	.070	.104	.084	.086
Sample 3	2839.79	2449.19	2282.7	1165	1115	1114	.00	.76	.81	.83	.063	.056	.054	.066	.069	.064
Sample 4	2552.45	2253.25	2185.1	1165	1115	1114	.00	.62	.69	.70	.08	.074	.072	.102	.083	.083
Sample 5	3468.57	2860.78	2700.6	1165	1115	1114	.00	.78	.83	.85	.051	.045	.043	.066	.047	.047
Sample 6	3523.03	2839.48	2492.9	1165	1115	1114	.00	.69	.77	.82	.081	.071	.063	.101	.080	.088
Sample 7	2481.53	2276.29	2049.4	1165	1115	1114	.00	.66	.74	.78	.057	.051	.043	.074	.061	.043
Mean								.70	.76	.78	.068	.062	.058	.086	.071	.069

Table 2. Factor correlations of alternative models applied to 7 datasets

	E~A	E~C	E~S	E~O	A~C	A~S	A~O	C~S	C~O	S~O
Observed scale scores										
Sample 1	.23	.13	.25	.30	.32	.11	.27	.27	.44	.22
Sample 2	.12	.03	.04	.24	.23	-.04	.18	.22	.19	.24
Sample 3	.22	.07	.34	.30	.26	.21	.29	.21	.25	.28
Sample 4	.29	.17	.16	.31	.30	.09	.34	.25	.25	.18
Sample 5	.30	.28	.32	.48	.31	.26	.29	.50	.41	.35
Sample 6	.17	.00	.28	.22	.23	.02	.23	.07	.07	.02
Sample 7	.60	-.09	.25	.22	-.03	.05	.19	.06	-.39	.01
Mean	.28	.08	.23	.30	.23	.10	.26	.23	.17	.19
SD	.16	.12	.10	.09	.12	.10	.06	.15	.28	.13
Simple Oblique CFA with no method Factor Model										
Sample 1	.27	.16	.27	.42	.36	.10	.32	.30	.51	.30
Sample 2	.17	.04	.01	.33	.20	-.05	.29	.29	.25	.23
Sample 3	.24	.08	.38	.39	.27	.22	.35	.26	.30	.40
Sample 4	.39	.24	.22	.51	.32	.12	.44	.30	.36	.20
Sample 5	.44	.35	.38	.60	.40	.35	.40	.60	.60	.44
Sample 6	.24	.02	.26	.33	.24	.03	.30	.07	.16	-.06
Sample 7	.43	-.02	.23	.24	.25	.17	.51	.12	.00	.15
Mean	.31	.12	.25	.40	.29	.13	.37	.28	.31	.23
SD	.11	.13	.12	.12	.07	.13	.08	.17	.20	.17
CFA with 1 Method factor Model										
Sample 1	.18	.04	.19	.41	.09	-.28	.07	-.03	.22	-.01
Sample 2	-.43	-.14	-.02	.11	-.00	-.10	-.15	.28	.12	.28
Sample 3	.18	-.02	-.11	.28	.23	-.18	.30	-.18	.20	.12
Sample 4	.13	-.05	-.03	.37	-.09	-.36	.24	-.12	.17	.08
Sample 5	.03	-.18	-.12	.37	-.16	-.27	-.02	.17	.22	-.01
Sample 6	.21	.02	.38	.26	.24	.06	.27	.08	.14	.14
Sample 7	.44	-.06	.21	.18	.13	.10	.25	.10	-.17	.08
Mean	.11	-.06	.07	.28	.06	-.15	.14	.04	.13	.10
SD	.27	.27	.08	.19	.11	.15	.18	.17	.16	.14
Mp and Mn Method Factors Model										
Sample 1	.19	.03	.19	.41	-.28	-.72	-.18	-.21	.12	-.13
Sample 2	-.40	-.19	-.07	.06	-.02	-.14	-.14	.27	.10	.26
Sample 3	-.32	-.25	.05	.12	-.08	-.51	-.08	-.13	.05	.09
Sample 4	.10	-.10	-.09	.31	-.09	-.37	.18	-.19	.06	.06
Sample 5	-.06	-.13	-.09	.40	-.27	-.42	-.10	.24	.26	.12
Sample 6	.08	-.12	.22	.21	.13	-.37	.19	-.23	.07	.06
Sample 7	.41	-.10	.18	.20	.06	.06	.16	.11	-.24	.03
Mean	.00	-.12	.06	.24	-.08	-.35	.00	-.02	.06	.07
SD	.27	.28	.09	.14	.13	.15	.25	.16	.22	.15

Note: Big Five factor correlations are labeled as: EA= Extraversion-Agreeableness; EC= Extraversion-conscientiousness; ES = Extraversion-Emotional stability; EO = Extraversion-Openness; AC = Agreeableness-conscientiousness; AS = Agreeableness – emotional stability; AO = Agreeableness – Openness; CS = Conscientiousness-Emotional stability; CO = Conscientiousness – Openness; SO = Emotional stability - Openness

Table 3. Reliability estimates of variables in alternative models applied to 7 datasets

	Extroversion				Agreeableness				Conscientiousness				Emotional Stability				Openness to experience			
	Ob	NoM	M	PN	Ob	NoM	M	PN	Ob	NoM	M	PN	Ob	NoM	M	PN	Ob	NoM	M	PN
Sample 1	.90	.94	.94	.94	.81	.91	.95	.93	.84	.94	.93	.93	.89	.89	.89	.90	.75	.93	.96	.95
Sample 2	.86	.93	.89	.89	.81	.96	.96	.96	.82	.95	.95	.95	.85	.90	.90	.90	.78	.96	.95	.95
Sample 3	.89	.95	.94	.92	.84	.95	.95	.91	.84	.94	.93	.93	.86	.89	.11	.83	.80	.94	.93	.92
Sample 4	.85	.92	.91	.91	.82	.98	.96	.96	.79	.86	.77	.80	.83	.85	.71	.66	.81	.95	.92	.92
Sample 5	.82	.94	.90	.91	.70	.96	.97	.95	.71	.92	.93	.92	.83	.90	.89	.92	.73	.91	.94	.95
Sample 6	.91	.94	.95	.93	.87	.97	.97	.95	.85	.91	.92	.96	.88	.93	.92	.74	.75	.90	.94	.94
Sample 7	.74	.88	.89	.87	.67	.94	.88	.96	.79	.97	.98	.98	.80	.84	.83	.84	.69	.92	.91	.86
Mean	.85	.93	.92	.91	.79	.96	.95	.95	.80	.93	.94	.92	.85	.89	.78	.83	.76	.93	.94	.93
SD	.06	.02	.03	.02	.07	.02	.03	.02	.05	.04	.02	.06	.03	.03	.30	.10	.04	.02	.02	.03

Note: Ob = Observed variable; NoM = No Method latent variable; M = Method latent variable; PN = Item wording factor latent variable

Table 4. Average Variance Components Explained by Trait, Method, and Error by CFA Models

	Model 1: No M		Model 2: 1-M Model			Model 3: 2-M Model		
Study	T*	E	T	M	E	T	Mp+Mn[2]	E
Sample 1	.365	.635	.285	.107	.608	.246	.152	.602
Sample 2	.337	.663	.273	.097	.630	.268	.117	.615
Sample 3	.376	.624	.280	.120	.600	.247	.170	.583
Sample 4	.333	.667	.250	.118	.632	.240	.146	.614
Sample 5	.284	.716	.165	.141	.694	.162	.155	.683
Sample 6	.393	.607	.380	.066	.554	.310	.177	.513
Sample 7	.242	.758	.212	.060	.728	.190	.073	.737
Mean	.333	.667	.264	.101	.635	.238	.141	.621

Note: * T = Trait; E = Error; M = Method; Mp+Mn = Method – positively and negatively worded items.

[2] Since Mp and Mn influence different IPIP items, we decided not to report them separately, since differences in variance might be due to differences in items as indicators of and/or item loadings of Mp and Mn.

Figure 1. CFA model of the IPIP with No Method Factor Estimated

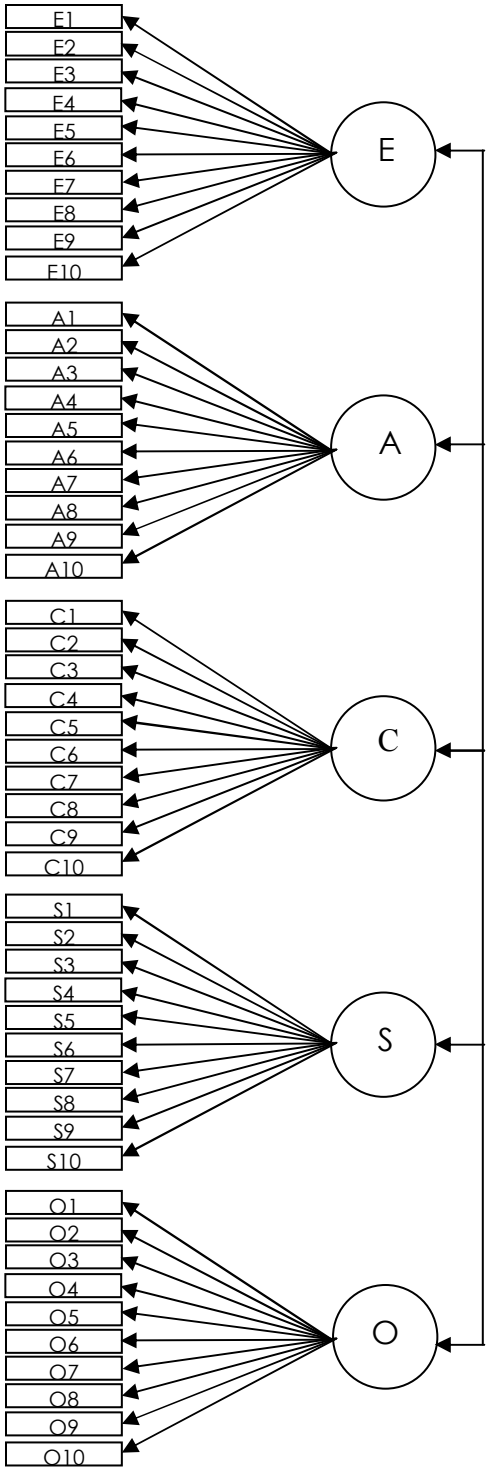

Figure 2. CFA Model of the IPIP with a Method Factor estimated

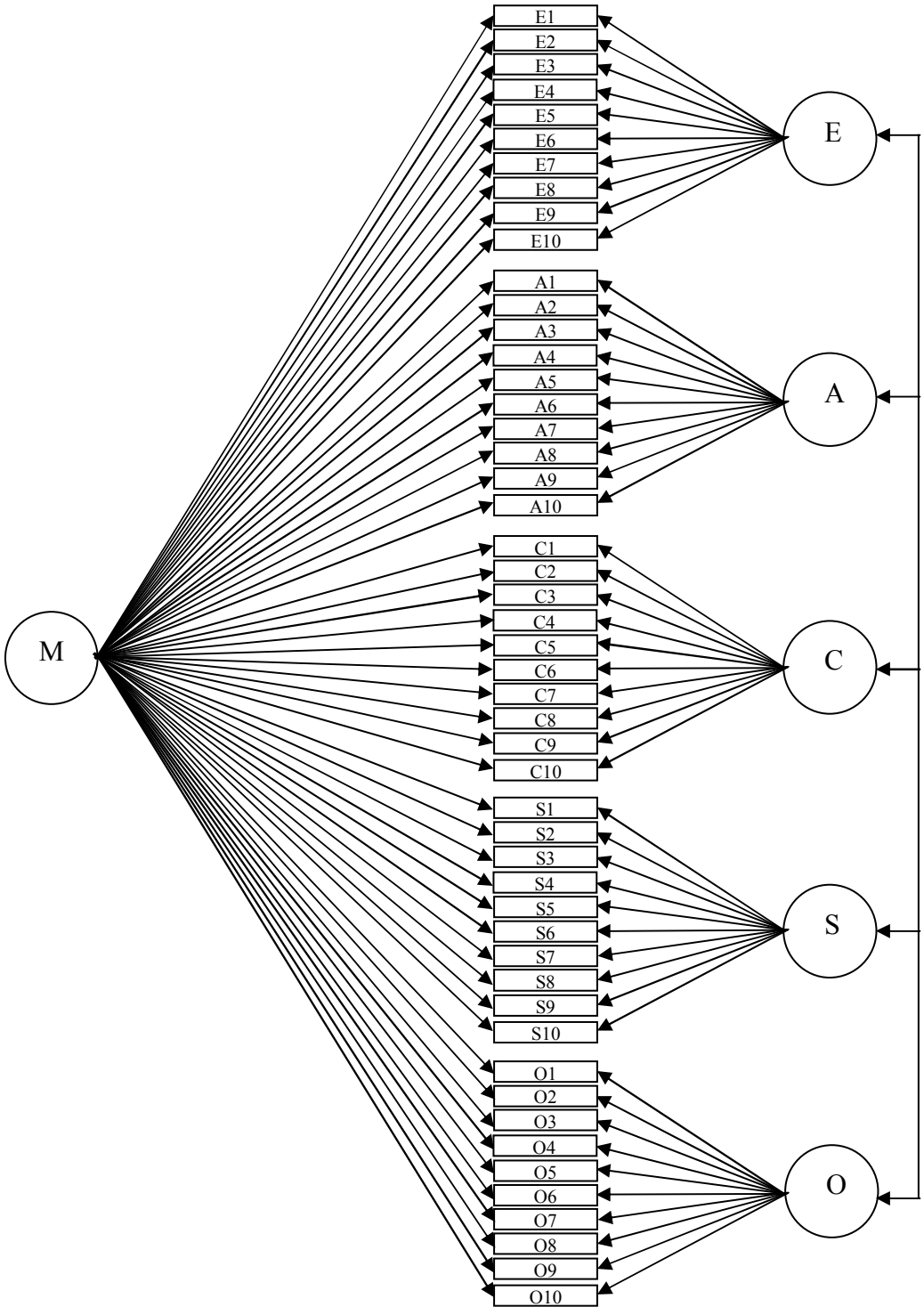

Figure 3. CFA Model of the IPIP with Positively and Negatively Worded Factors Estimated