The ubiquity of common method variance: The case of the Big Five

Michael D. Biderman [a,*], Nhung T. Nguyen [b], Christopher J.L. Cunningham [a], Nima Ghorbani [c]

[a] Department of Psychology, The University of Tennessee at Chattanooga, United States
[b] Department of Management, Towson University, United States
[c] Department of Psychology, University of Tehran, Iran

ARTICLE INFO

ABSTRACT

The factor structures of the International Personality Item Pool (IPIP) and NEO-FFI Big Five questionnaires were examined via confirmatory factor analyses. Analyses of IPIP data for five samples and NEO data for one sample showed that a CFA model with three method bias factors, one influencing all items, one influencing negatively worded items, and one influencing positively worded items fit the data significantly better than models without method factors or models with only one method factor. With the method factors estimated, our results indicated that the Big Five dimensions may be more nearly orthogonal than previously demonstrated. Implications of the presence of method variance in Big Five scales are discussed.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

In the past 30 years there has been resurgence in the study of personality in psychology due primarily to the discovery of a common factor structure underlying measures of personality characteristics. The dominant taxonomy is a lexically based five-factor structure originally developed within countries that use Northern European languages (e.g., Saucier & Goldberg, 2003). Most popularly known as the Big Five, this framework includes the traits of Extraversion (E), Agreeableness (A), Conscientiousness (C), Emotional Stability (S, often measured as Neuroticism), and Openness to Experience (O, sometimes measured as Intellect).

Despite the wide acceptance and application of this personality framework, several measurement-related issues have continued to challenge personality researchers. In particular, although conceived of as orthogonal dimensions of personality, correlations between summated scale scores on most Big Five personality tests are generally moderately positive (e.g., Digman, 1997; Mount, Barrick, Scullen, & Rounds, 2005). There are at least two explanations for this. The first is that the five factors commonly estimated are actually themselves correlated and perhaps indicators of higher order factors. More specifically, it has been suggested that the Big Five factors are indicators of the higher order factors of *stability* (as indicated by agreeableness, conscientiousness, and the inverse of neuroticism) and *plasticity* (as indicated by openness and extra-

version) (DeYoung, Peterson, & Higgins, 2001; Digman, 1997). Others have alternatively suggested that there may be one overriding personality factor, deemed *evaluation* (Goldberg & Somer, 2000; Saucier, 1997), the "Big One" (Musek, 2007) or the general factor of personality (GFP) (Rushton, Bons, & Hur, 2008; Van der Linden, Nijenhuis, & Bakker, 2010).

A second explanation for the commonly identified positive relationships among Big Five scale scores is that there is a separate source of influence that affects responses to all items in these questionnaires, and that this influence is somehow distinct from that of the Big Five factors themselves. Often this type of shared influence across scores collected using a specific method is referred to as common method bias (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). The word bias in this context refers to the individual differences that become manifest when the same method is used across multiple personality scales. Associated with this common bias is the notion of common method variance, which, in the present context, can be understood as variance in Big Five scale item responses throughout a measure that is due to the influence of common method bias.

The existence of common method variance has been recognized in questionnaire research for many years (e.g., Cote & Buckley, 1987; Doty & Glick, 1998). Most research on this topic has been based on analyses of multitrait–multimethod data from a single measure, usually an isolated scale or domain score, per trait-method combination. Although helpful in highlighting the potential impact of common method bias, such investigations have not made it possible to separate within-dimension covariance from between-dimension covariance (Tomas, Hontangas, & Oliver, 2000). Indeed, until recently, the study of common method variance based on

* Corresponding author. Address: Department of Psychology, The University of Tennessee at Chattanooga, 615 McCallie Avenue, Chattanooga, TN 37403, United States.
    E-mail address: Michael-Biderman@utc.edu (M.D. Biderman).

analyses of individual items or representative item parcels has been neglected. This is unfortunate, given that such analyses are necessary to properly estimate and compare within- and between-dimension variability (Marsh, Scalas, & Nagengast, 2010).

One of the first studies to permit this type of variability separation using Big Five measure data was Schmit and Ryan (1993), in which analyses of multiple item composites from each dimension revealed the potential for measures of the Big Five traits to include common method variance. Schmit and Ryan factor analyzed responses to item composites of the NEO-FFI (Costa & McCrae, 1989) within a work context using applicant and non-applicant samples. An exploratory factor analysis (EFA) of the 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" (Schmit & 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.

Beginning in the late 1990s confirmatory factor analyses (CFAs) were conducted of questionnaire items or parcels to identify and study method biases. These studies included analyses of data collected with the Rosenberg Self-Esteem scale (Marsh, 1996; Marsh et al., 2010; Motl & DeStefano, 2002; Tomás & Oliver, 1999) and investigations of the possibility that common method bias may represent or reflect respondent faking or socially desirable responding in certain situations (e.g., Biderman & Nguyen, 2004; Bäckström, 2007; Bäckström, Björklund, & Larsson, 2009; Cellar, Miller, Doverspike, & Klawsky, 1996; Klehe et al., 2011; Ziegler & Buehner, 2009). In these experimental studies, participants were typically asked to respond to Big Five measure items under faking and no-faking conditions. In the faking conditions of these studies, variance common to all items was represented by a single latent variable similar to what Podsakoff et al. (2003) labeled an "unmeasured method" effect. However, with the exception of Bäckström (2007) and Bäckström et al. (2009) the study of common method variance in Big Five questionnaires in nonapplicant honest conditions has received little attention. This is problematic, given that in most scenarios, participants are instructed to do precisely that – respond honestly.

This limitation of previous research in this area combined with the fact that most personality assessment is by self-reported completion of personality inventories, leaves a major deficit in our understanding of what is actually being assessed when we use common personality measures such as those designed to capture the Big Five traits. Further complicating matters is a common recommendation for developing and/or choosing assessment items for self-report measures, that encourages the inclusion of both positively worded items ("I am the life of the party") and negatively worded items (e.g., "I don't talk a lot") in a single scale. The logic behind including both types of items is that their presence might reduce the effects of participant response tendencies such as acquiescence (DeVellis, 1991; Nunnally, 1978; Spector, 1998). This recommendation has been so widely shared that the practice of using negatively worded items to presumably counteract respondents' acquiescence can be found throughout most areas of organizational research including personality assessment (e.g., Paulhus, 1991), leadership behavior (e.g., Schriesheim & Eisenbach, 1995; Schriesheim & Hill, 1981), role stress (Rizzo, House, & Lirtzman, 1970), job characteristics (Harvey, Billings, & Nilan, 1985), and organizational commitment (e.g., Meyer & Allen, 1984).

Unfortunately, the negatively worded items that were introduced to counter individuals' response tendencies have been found to increase systematic and perhaps construct-irrelevant variance in scale scores in studies: (a) of self-esteem (e.g., Hensley & Roberts,

1976; Marsh, 1996; Marsh et al., 2010; Motl & DeStefano, 2002; Tomás & Oliver, 1999), (b) using Rizzo, House, and Lirtzman's (1970) role conflict and role ambiguity scale (McGee, Ferguson, & Seers, 1989), (c) using Meyer and Allen's (1984) Organizational Commitment scale (Magazine, Williams, & Williams, 1996), (d) using Spector's (1988) Work Locus of Control Scale, and (e) using Hackman and Oldham's (1975) Job Diagnostic Survey (Idaszak & Drasgow, 1987). In addition to increased "noise" interjected by such items and the potential multidimensionality introduced by negatively worded items, the inclusion of such items in leadership behavior measures has been shown to decrease a scale's reliability and validity (Schriesheim & Eisenbach, 1995; Schriesheim & Hill, 1981).

Recently, Marsh et al. (2010), using confirmatory factor analyses, provided evidence for two conclusions regarding the factorial structure of questionnaires employing negatively worded items. First, Marsh et al. found that a model with two method factors (one influencing only positively worded items and the other influencing only negatively worded items) fit the data of the Rosenberg Self Esteem (RSE) scale better than models without method factors and better than models with only one wording-type factor. Although other researchers had found item wording influences associated with negatively worded items (e.g., DiStefano & Motl, 2006), Marsh et al.'s results provided evidence for analogous influences associated with positively worded items. Marsh et al. further found, based on longitudinal models, that positive and negative influences were not sporadic and spontaneous, but substantive and stable over time. These two findings coupled with other studies in which method factors have been implicated (Cote & Buckley, 1987; Doty & Glick, 1998) suggest that method effects including item-wording specific method effects may be influential whenever personality is assessed using self-report questionnaires.

Given the mounting evidence for the prevalence of common method variance in personality assessment and the increasing usage of personality assessments in organizational research and practice, it is surprising that few attempts have been made to examine the effects of method bias and item wording biases on the factor structure of Big Five measures. As mentioned previously, studies estimating a common method factor have for the most part focused on identifying socially desirable responding. Apart from those above-mentioned studies, there have been no published CFA models of Big Five questionnaire data that have included item-wording factors. For all the reasons already stated, the main purpose of the present study was to closely examine the factor structures of two commonly used Big Five questionnaires, the IPIP and NEO-FFI with the intent of assessing the extent to which responses to items in these questionnaires are influenced by method factors and/or wording-specific method factors. This was done in a fashion similar to that used by Marsh et al. (2010), by comparing CFA models with different assumptions concerning general method factors and wording-specific method factors.

The specific models that were compared in this study are presented in Fig. 1. Within this figure, Model 1 is a basic CFA of a Big Five questionnaire with correlated trait factors but no method factor. If there were a common method influence on all 50 items of this instrument, the presence of such an influence would have to be accommodated in the model by increased positive correlations among the factors (Paglis & William, 1996; Williams & Brown, 1994). In Model 2, a single method factor, M, has been added to the basic CFA of Model 1 (e.g., Bäckström, 2007; Bäckström et al., 2009; Cellar et al., 1996). In this model, M is defined as an "unmeasured" method factor in that it has no unique indicators but rather is estimated from indicators of the Big Five factors. M is a first order factor whose indicators are items which also are indicators of the Big Five factors, not a higher order factor. This type of model
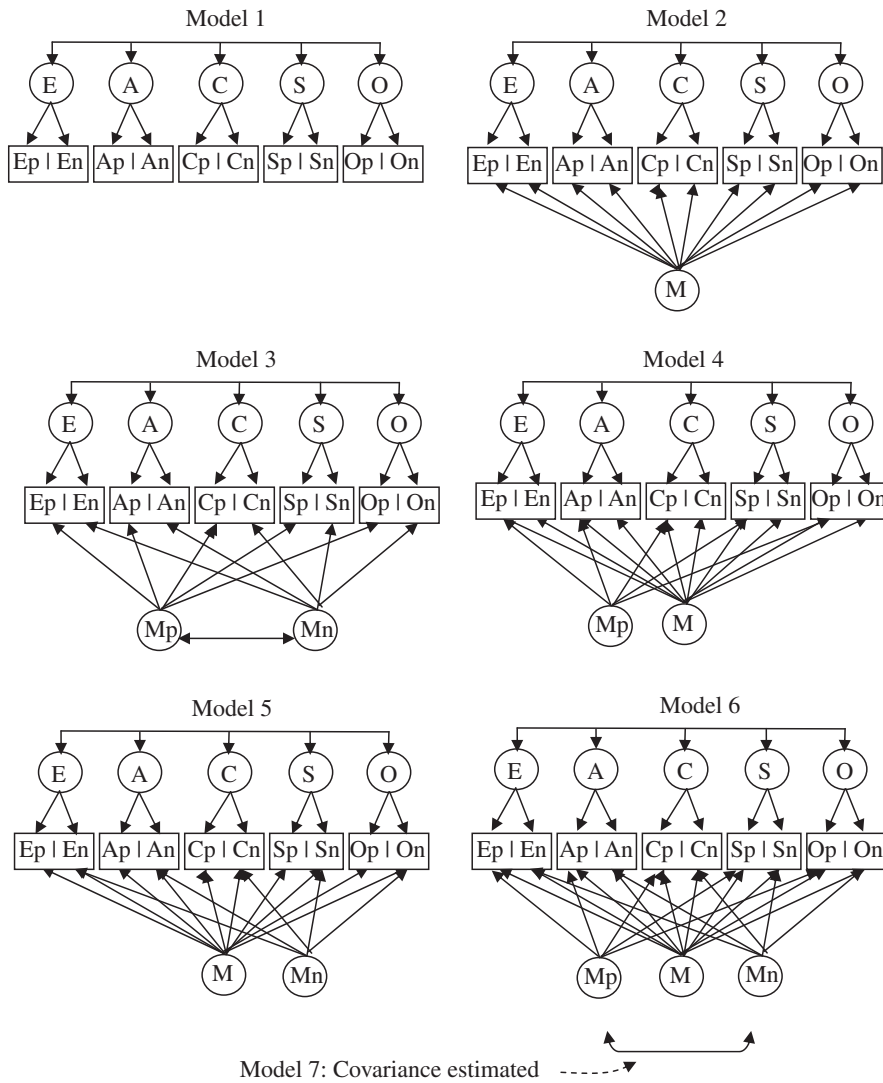
Fig. 1. Models compared. Each rectangle represents the items indicating a Big Five dimension. The left half of each rectangle represents positively worded items and the right half negatively worded items. A single arrow drawn from a factor to a rectangle represents all the loadings of the indicators represented by the rectangle on that factor. Residual latent variables have been omitted for clarity.

in which observed variables indicate multiple factors has been called a bifactor model (e.g., Chen, West, & Sousa, 2006).

Model 3 is a model with two separate first order method factors – one influencing only positively worded items and the other influencing only negatively worded items. This is analogous to the model found to best represent the RSE data previously mentioned. In that study, Marsh et al. (2010) had to restrict the two method factors to orthogonality due to the fact that the RSE assesses only one trait dimension. Because Big Five questionnaires assess five trait dimensions, it is possible to apply a slightly more elaborate model than that of Marsh et al.'s, estimating the covariance between the two item-wording factors. Note that Model 2 is a special case of Model 3 in which the correlation between the two separate item-wording factors is set to one.

Models 4–6 are generalizations of Model 2 with both an overall common method factor *and* one or more method factors specific to a wording type – either a positive wording method factor in Model 4 or a negative wording factor in Model 5, or both in Model 6. In these models, the correlations between method factors are restricted to zero. Because the method factors were restricted to orthogonality in these models, these models are not generalizations of Model 3 although they are generalizations of Model 2. Be-

cause of this, it is not possible to use chi-square difference tests to compare Models 4–6 with Model 3, although they can be compared using this test with Model 2. The last model in the series compared here, Model 7, was a simple generalization of Model 6 in which the correlation between Mp and Mn was estimated. Only Mp and Mn were allowed to covary in Model 7. Mp and Mn were constrained to be orthogonal to M. Model 7 was a generalization of Model 3 and thus chi-square difference tests could be used to compare the fit of Model 7 with that of Model 3.

To be clear, the conceptualization of Models 3 vs. 4 through 7 are quite different. For Model 3, each of the two method biases influence only one type of item – Mp influences only positively worded items and Mn influences only negatively worded items. This acknowledges the distinction between positive and negative wordings, but does not account for any general influence that might operate on all the items, such as faking or socially desirable responding. Models 4 through 7, however, include a general influence while allowing for a second, wording specific influence on either positively worded or negatively worded items or both. Thus, if there are individual differences in a personal characteristic affecting all the items, such as socially desirable responding, for example, those would be represented by M. On the other hand,

individual differences specific to item wording would be represented by either Mp or Mn.

In addition to investigating the need to incorporate method bias factors in the analysis of Big Five questionnaire data, we addressed an implication of the presence of method bias for estimated relationships involving Big Five dimensions. Specifically, we investigated the extent to which Big Five dimensions are correlated after common method variance is taken into account. Whether the resulting correlations are essentially zero or even negative is a key issue for higher order factor theories of the Big Five (Digman, 1997; Musek, 2007). If the dimensions remain positively correlated, this leaves open the possibility that higher order factors could account for those correlations. Alternatively, uncorrelated Big Five dimensions after accounting for method variance would be evidence against the possibility that there are substantively meaningful higher order dimensions of which the Big Five factors are indicators.

Based on the above discussion, we entertained the following hypotheses.

H1: Adding a common method factor will significantly improve the measurement model fit.
H2: Models with separate item wording method factors will have significantly better fit than a model estimating only one method factor.

Although the results of studies of socially desirable responding suggest that common variance is increased when instructions or incentives to fake are present, we have no evidence that such an increase is due to the influence of a single bias factor, M, or to the joint influences of both Mp and Mn. For this reason, we had no specific reason to expect differences between Model 3 vs. Models 4 through 7 so no hypotheses concerning differences in fit between them are presented and the analyses comparing them are treated as exploratory.

As argued above, the presence of interitem correlations would affect estimates of correlations between the Big Five latent variables in models without common method factors. Failure to account for influences of such factors would positively bias estimates of correlations between the Big Five dimensions. As it is also not possible to account for method factors when correlating scale scores, we would expect correlations between scale scores to be similarly positively biased. Accounting for interitem correlations with method factors, however, should reduce the positive bias in estimates of the Big Five dimensions. For this reason we propose the following hypothesis.

H3: Correlations between factors in models that include common method factors will be less positive than the correlations between raw scale scores.

On the assumption that H3 would be supported, we explored the extent to which the estimates of Big Five dimensions changed when those estimates were obtained within the context of method bias models.

2. Method

2.1. Participants

Data for this research came from five separate samples.

2.1.1. Sample 1

Participants were 183 undergraduate students enrolled in undergraduate courses at a southeastern university in the United States including 82 males and 101 females. Participants were those from a larger sample who received only instructions to respond honestly before completing the questionnaire. Mean age was 24.34 (SD = 13.37). Ethnicity was 69.9% White, 21.3% African American, and 8.8% "other" (Damron, 2004). Participants were given the Wonderlic Personnel Test (WPT: Wonderlic, 1999) prior to being given the Big Five questionnaire.

2.1.2. Sample 2

Participants were 202 undergraduate students enrolled at a southeastern university in the United States with 49 males and 153 females. Mean age was 19.2 years (SD = 2.5). Ethnicity was 59.6% White, 34.7 African American, and 5.7% Other. (Biderman & Nguyen, 2009). Participants were given the WPT prior to being given the Big Five questionnaire.

2.1.3. Sample 3

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 behaviors. There were 111 males and 200 females. Mean age was about 21 years. All participation was voluntary for course credit and/or raffle drawing (Cunningham, 2007).

2.1.4. Sample 4

Participants were 404 undergraduates enrolled at the University of Tehran with 148 males and 256 females. Participation 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 21.53 (SD = 1.19).

2.1.5. Sample 5

Participants were 189 undergraduates and graduate students enrolled at a southeastern university in the USA with 61 males and 128 females and two respondents whose gender was not reported. Seventy percent were White, 23% African American, and 7% Others. Mean age was 21.60 (SD = 5.33). Participation was voluntary in exchange for course credit (Biderman, Nguyen, & Cunningham, 2009).

2.2. Measures

2.2.1. Big Five personality traits

The 50-item sample questionnaire from the IPIP web site (http://ipip.ori.org/ipip/; Goldberg, 1999) was used in all samples. For samples 1–5, the order of items was identical to that of the Sample 50 item scale provided on the IPIP web site: EACSO repeated 10 times. For sample 3 the order of dimensions was the same except that all items within each dimension were presented together before items from the next dimension. For Sample 4, 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.

2.2.2. Neo

In sample 5, the 60-item NEO-FFI (Costa & McCrae, 1989) was administered in addition to the IPIP, with about half of the participants receiving the IPIP first. The order of items was SEOCA repeated 12 times. The NEO responses were entered in a rectangular response area that allowed the entry of five responses per line. A similar-appearing response block was created for the IPIP questionnaire used in this study.

In all samples, participants were instructed to respond honestly to each personality item. Response anchors for all items ranged from "1" = very inaccurate to "5" = very accurate except for dataset 2 in which the response scale ranged from "1" = very inaccurate to "7" = very accurate. Internal consistency estimates for all scale scores are summarized in Table 1.

2.3. Analyses

A series of CFA models was applied to the data of each sample. Because of the focus on item-wording differences, factor indicators in each model were individual items. Negatively worded items for all but the neuroticism dimension were reverse-coded; for the neuroticism dimension, the neuroticism items from the NEO-FFI were reverse-coded so that higher scores represented a higher degree of emotional stability or the S dimension of the IPIP questionnaire. All models were estimated using M*plus* V5.2 (Muthén & Muthén, 1998–2007). Maximum likelihood estimates were obtained for all models. Some analyses were conducted using factor scores as in Bollen and Paxton (1998). The factor scores were generated using the regression method (Muthén, 1998–2004) and were imported into a common statistical package for computation of correlations necessary for testing hypotheses. The seven models shown in Fig. 1 were applied to each dataset.

Mode1 was a straightforward CFA of the Big Five items. Model 2 was identical to Model 1 except that a sixth, first order latent variable, M, was included. All items were required to load on M. For

Table 1
Means, standard deviations and correlations between Big Five scales The rightmost entry in each line is the scale reliability.

Analysis	Mean	SD	1	2	3	4	5
1. Extraversion							
1	3.380	0.841	.896				
2	4.507	0.974	.845				
3	3.347	0.871	.905				
4	2.156	0.697	.742				
5	3.373	0.744	.892				
6	3.547	0.520	.777				
2. Agreeableness							
1	4.064	0.596	.198	.823			
2	5.250	0.822	.275	.788			
3	4.032	0.639	.165	.865			
4	2.875	0.540	.302	.673			
5	4.063	0.499	.308	.815			
6	3.659	0.551	.388	.794			
3. Conscientiousness							
1	3.614	0.604	.064	.268	.791		
2	4.793	0.805	.035	.347	.778		
3	3.571	0.714	.003	.230	.853		
4	2.570	0.733	−.010	.267	.787		
5	3.595	0.611	−.006	.288	.839		
6	3.474	0.269	.139	.027	.845		
4. Stability							
1	3.171	0.737	.229	.141	.221	.836	
2	3.979	1.095	.207	.163	.057	.861	
3	3.307	0.801	.284	.022	.069	.882	
4	1.904	0.771	.181	.131	.100	.803	
5	3.158	0.751	.293	.198	.189	.888	
6	3.269	0.744	.386	.296	.081	.889	
5. Openness							
1	3.669	0.542	.224	.241	.126	.183	.775
2	4.727	0.730	.159	.276	.165	.226	.766
3	3.507	0.556	.223	.227	.069	.021	.748
4	2.628	0.569	.173	.325	.112	.091	.692
5	3.748	0.495	.062	.073	.047	−.005	.772
6	3.233	0.531	−.087	−.083	−.299	−.035	.774

Note. Analyses 1–5 represent the IPIP questionnaire from samples 1 through 5. Analysis 6 represents the NEO questionnaire from sample 5. The NEO Neuroticism scale was reverse scored to make it comparable with the IPIP stability scale.

purposes of model identification, M was constrained so that it was orthogonal to all of the Big Five factors (Williams, Ford, & Nguyen, 2002). Thus, M represented variance shared among all 50 items that was not accounted for by the Big Five factors.

Model 3 was a generalization of Model 2 with two factors: one indicated only by positively worded items (Mp) and the other by negatively worded items (Mn). Both Mp and Mn were constrained to be orthogonal to the Big Five latent variables, although the correlation between Mp and Mn was estimated. For the IPIP questionnaire, there were 26 positively worded items, with 5, 6, 6, 2, and 7 items from the E, A, C, S, and O Big Five dimensions, respectively. For the NEO-FFI questionnaire, there were 12 items per dimension, and there were 29 positively worded items, with 8, 4, 8, 4, and 5 positive items indicating the E, A, C, S, and O Big Five dimensions respectively. Models 4–6 included the overall method factor of Model 2 with the item-wording factors of Model 3 estimated as orthogonal to the overall method factor. Model 7 was a generalization of Model 6 in which the covariance between Mp and Mn was estimated.

We used four goodness-of-fit statistics for model evaluation: 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 (i.e., measurement model misspecification), SRMR was found to be most sensitive to misspecified factor covariances (i.e., 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 reported as 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).

3. Results

Space limitations prevent the presentation of covariance matrices of all individual items, although those are available from the first author. To provide some indication of the comparability of these five datasets with others, Table 1 presents correlations of Big Five scale scores for all six administrations of questionnaires – IPIP questionnaires from all five samples and the NEO from one sample – along with means, standard deviations, and reliability coefficients for these scales. Inspection of the table shows that most of the inter-scale correlations were positive. Mean inter-scale correlations for the six correlation matrices were .19, .10, .13, .17, .14 for the IPIP scale scores, and .08 for the NEO-FFI scale scores.

3.1. Goodness of fit

The following summarizes the results of six analyses for each model – those of the IPIP questionnaire in five samples and the NEO-FFI questionnaire administered to sample 5. Thus analyses 5 and 6 were both obtained using sample 5 data. The first seven lines of Table 2 present the fit statistics of Models 1 through 7 shown in Fig. 1 – the models relevant to Hypotheses 1 and 2. Chi-square difference tests were performed to compare goodness-of-fit of the models. As shown in line 9 of Table 2, across all six analyses Model 2 had a significantly better fit to the data than Model 1, with $\Delta\chi^2$ ranging from 272.7 to 672.9, $p < .001$ for each. CFI, RMSEA, and SRMR values are presented in Table 3. Across the six analyses (five of IPIP data and one of NEO data), the CFIs from Model 2 were larger than those of Model 1. Moreover, both RMSEA and SRMR values

Table 2
Chi-square goodness-of-fit measures and chi-square difference tests.

		Analysis						
		1	2	3	4	5	6	df
Questionnaire:		IPIP	IPIP	IPIP	IPIP	IPIP	NEO	IPIP/NEO
	Model 1	2174.4[b]	2461.3[b]	3523.0[b]	3734.4[b]	2529.6[b]	3214.8[b]	1165/1700
	Model 2	1901.7v	2155.8[b]	2839.5[b]	2431.2[b]	2210.5[b]	2903.9[b]	1115/1640
	Model 3	1853.7[b]	2092.8[b]	2492.9[b]	2275.0[b]	2177.7[b]	2851.8[b]	1114/1639
	Model 4	1786.0[b]	1993.5[b]	2375.2[b]	2112.9[b]	2050.3[b]	2742.7[b]	1089/1611
	Model 5	1758.2[b]	2024.5[b]	2389.5[b]	2152.7[b]	2044.1[b]	2732.7[b]	1091/1609
	Model 6	1642.9[b]	1883.4[b]	2148.9[b]	1894.6[b]	1895.2[b]	2594.3[b]	1065/1580
	Model 7	1619.6[b]	1841.6[b]	2139.1[b]	1853.2[b]	1887.6[b]	2592.1[b]	1064/1579
	Model 7 r_{MpMn}	.70	.68	.25	.49	.27	.20	
$\Delta\chi^2$	Model 2 vs. 1	272.7[b]	299.2[b]	459.3[b]	672.9[b]	327.1[b]	326.5[b]	50/60
$\Delta\chi^2$	Model 7 vs. 2	282.1[b]	314.2[b]	700.4[b]	578.0[b]	322.9[b]	311.8[b]	51/61
$\Delta\chi^2$	Model 7vs. 3	234.1[b]	251.2[b]	353.8[b]	421.8[b]	290.1[b]	259.7[b]	50/60
$\Delta\chi^2$	Model 7 vs. 6	23.2[b]	41.8[b]	9.8[b]	41.4[b]	7.6[a]	2.2	1/1

Note. For analysis 4, residual variance of one item was set to .001.
[a] $p < .01$.
[b] $p < .001$.

Table 3
CFI, RMSEA, and SRMR goodness-of-fit statistics.

		Analysis					
		1	2	3	4	5	6
Questionnaire:		IPIP	IPIP	IPIP	IPIP	IPIP	NEO
CFI	Model 1	.696	.644	.686	.663	.664	.631
	Model 2	.763	.715	.771	.742	.730	.692
	Model 3	.777	.732	.817	.783	.738	.704
	Model 4	.790	.752	.829	.789	.763	.724
	Model 5	.799	.744	.827	.786	.765	.726
	Model 6	.826	.776	.856	.816	.795	.753
	Model 7	.833	.787	.857	.825	.797	.753
RMSEA	Model 1	.069	.074	.081	.057	.079	.069
	Model 2	.062	.068	.071	.051	.072	.064
	Model 3	.060	.066	.063	.047	.071	.063
	Model 4	.059	.064	.062	.046	.068	.061
	Model 5	.058	.065	.062	.047	.068	.061
	Model 6	.054	.062	.057	.044	.064	.058
	Model 7	.053	.060	.057	.043	.064	.058
SRMR	Model 1	.099	.096	.101	.074	.101	.092
	Model 2	.079	.079	.079	.060	.081	.075
	Model 3	.080	.081	.088	.057	.083	.076
	Model 4	.075	.077	.076	.056	.079	.073
	Model 5	.075	.075	.085	.056	.074	.071
	Model 6	.073	.074	.064	.055	.072	.069
	Model 7	.068	.070	.064	.050	.072	.068

consistently indicated better fit for Model 2 than Model 1. Taken together, these fit indices indicated that including only a single common method factor improved fit of the CFAs to Big Five data across all samples. There were no large differences between the five analyses of IPIP data and the analysis of NEO data. Hypothesis 1 was fully supported.

Hypothesis 2 stated that estimating two method factors – one represented by positively worded items and one by negatively worded items would significantly improve the CFA model fit. This hypothesis was tested by comparing the goodness of fit of Model 2, with just one general method factor, with that of Model 7, with a general method factor and two addition wording specific factors, Mp and Mn. Since Model 2 is a special case of Model 7 with Mp and Mn loadings set equal to 0 and the correlation of Mp with Mn fixed, the chi-square difference test comparing the fit of the two models was used. The results of the tests are reported in Line 10 of Table 2. These tests revealed that Model 7 fit significantly better in all comparisons, with chi-square differences ranging from 282.1 to 700.4. From Table 3, the CFI and the RMSEA fit indices also indicated better fit for Model 7 than Model 2 in every analysis.

These results indicate that the data across these multiple sets were better represented by including, in addition to a general method factor, a method factor for positively worded items to be separate from the factor for negatively worded items. Thus, Hypothesis 2 was fully supported.

Additional comparisons were made to ascertain the need for modeling all three method factors. These comparisons were of Model 7 with Model 3 to test the need for a general method factor when wording specific factors are already present. The results of these comparisons are in Line 11 of Table 2, with chi-square difference values ranging from 234.1 to 421.8, $p < .001$ for each. Finally, comparisons of Model 7 with Model 6 were made to determine whether Mp and Mn should be correlated in the context of a common method factor. These comparisons are presented in Line 12 of Table 2. Five of the six comparisons resulted in significant differences, with chi-square differences ranging from 7.6 to 41.8. One comparison, involving Analysis 6 (the NEO FFI data of sample 5) equaled 2.2 and did not reach traditional levels of statistical significance ($p > .05$).

Although the test statistic values are not presented in Table 2, comparisons of Models 4 and 5 with Model 6 resulted in significant chi-square difference statistics, suggesting that when a common method factor is estimated, failure to account for *both* positively worded items and negatively worded items with their own factors will yield significant decrements in goodness-of-fit. All of the above results taken together suggest that there may be three separate influences on responses to each item in a typical Big Five questionnaire: the influence of the personality dimension represented by the item, a general method bias, and an influence specific to the wording of the item (positive vs. negative wording).

Although Model 7 (allowing Mp and Mn to correlate) fit significantly better than Model 6 (requiring that Mp and Mn be orthogonal) for five of the six analyses, we encountered occasional anomalous results in the application of Model 7. These suggested that this model may be overparameterized for samples of the sizes used here. Since the differences in goodness-of-fit between Models 6 and 7 were small and since most of the remaining analyses did not require that Mp and Mn be correlated, with one exception we based the remainder of the results on estimates from application of Model 6.

3.2. Convergent validity of Model 6 latent variables across questionnaires

Participants in sample 5 responded to both the IPIP and the NEO instruments, making it possible to assess convergent validity of la-

Table 4
Mean trait and method factor loadings, mean proportion of item variance due to traits, and mean proportion of item variance due to method factors for Model 6.

Analysis	Mean loading on each factor Factor								Mean proportion of item variance due to traits	Mean proportion of item variance due to methods
	E	A	C	S	O	M	Mp	Mn		
IPIP questionnaire										
1	.620	.553	.475	.157	.404	.215	.172	.262	.251	.170
2	.549	.371	.476	.305	.421	.268	.090	.266	.216	.187
3	.643	.541	.467	.255	.456	−.066	.335	.374	.269	.226
4	.457	.345	.447	.050	.371	.160	.252	.122	.172	.129
5	.568	.500	.553	.262	.503	.261	.052	.179	.263	.184
NEO-FFI questionnaire										
6	.194	.231	.530	.489	.469	.265	−.017	.215	.199	.177

tent variables estimated from the application of Model 6 to the sample 5 data of each instrument. A dual version of Model 6 was created in which the model applied to the IPIP data and the model applied to the NEO data were connected by allowing the latent variables in the two models to be correlated. Correlations between corresponding latent variables from the two datasets were .93, .80, .96, .92, and 81 for E, A, C, S, and O respectively. Of potentially greater interest here are the correlations of .87, .79, and .58 for M, Mp, and Mn respectively, suggesting high convergent validity across Big Five instruments for these three method factors.

3.3. Proportion of method variance and its effect on correlations

Although the results of the chi-square difference statistics suggested that method bias factors significantly improved goodness-of-fit, they give little indication of the amount of variance of items accounted for by the various method factors. To this end, mean loadings on each factor were computed and are presented in Table 4. Moreover, the mean proportion of variance due to trait factors and the mean proportion of variance due to method factors are presented in the rightmost columns of the table. Since in this model all three factors affecting each item were orthogonal, the proportion of variance of an item due to traits is simply the square of the loading on its trait factor (e.g., Williams, Cote, & Buckley, 1989). Similarly, the proportion of variance of an item due method factors is the sum of its squared loadings on M and either Mp or Mn depending on the item wording. The values in Table 4 are the means of the proportions across items in each questionnaire. The mean proportion of explained variance of the items is simply the sum of the rightmost two values for each analysis. The proportion of explained variance due to method factors is the rightmost value in each line divided by the sum of rightmost two values. As shown in the table, the mean proportion of explained item variance due to method factors was substantial in each sample, just slightly lower than the proportion due to trait factors.

An alternative way of summarizing the effect of method bias is presented in Table 5. In this table, correlations of raw scale scores with factor scores from Model 6 are presented. In this table, the factor scores are considered to be the "purer" measures of each

dimension, uncontaminated by method effects, while the scale scores are contaminated to an extent determined by item loadings on the method factors. Inspection of this table shows that the correlations were generally large, with the exception of correlations with scale scores representing the S dimension for the IPIP questionnaire. For the NEO items, both E and A scale scores exhibited low correlations with factor scores. These results suggest that scale scores cannot always be counted on to give a veridical representation of the underlying trait dimension.

3.4. Orthogonality of Big Five dimensions

Hypothesis 3 stated that all Big Five factors will be less positively correlated when estimated from models with method factors than scale scores computed without method factors. Table 6 presents correlations between Big Five latent variables taken from the M*plus* output from application of Model 6 along with factor determinacies. Factor determinacies are estimates of the correlation of factor scores with the latent variables they represent and have been called factor validities by Grice (2001) who stated that values greater than .90 may be necessary if factor score estimates are to serve as adequate substitutes for the factors themselves. As can be seen from the table, most of the factor determinacy values are larger than .9 and only one, that for agreeableness from Analysis 4, is less than .8. Comparing the correlations in Table 6 with those between the Big Five scales in Table 1 shows that the correlations between latent variables are generally less positive than the correlations between scale scores. The mean of all the IPIP scale correlations from Table 1 was .15, with only 6 of the 50 correlations less than zero. On the other hand, the mean of all the IPIP latent variable correlations in Table 6 was .01, with 29 of 50 negative correlations. This pattern of results was similar for the NEO analysis, with the mean of the correlations decreasing from .08 to −.13. These results supported Hypothesis 3.

With respect to the possibility that the Big Five factors are indicators of higher order factors, Big Five dimensions A, C, and S have been proposed as indicators of the higher order factor labeled Stability (DeYoung et al., 2001) From Table 6, the mean correlation of A with C was .16, A with S was −.22, and C with S was −.13. Thus, two of the three mean correlations between indicators of the proposed stability higher order factor were negative. The mean correlation of E with O, indicators of the higher order factor of plasticity, was .09.

Although the mean of the correlations of Big Five latent variables across studies was close to zero for the IPIP analyses, the individual correlations varied about that mean. To determine whether the dimensions within each sample could be considered to be precisely orthogonal, a restricted version of Model 6 with all Big Five latent variable correlations set to zero was applied to each dataset. Chi-square difference statistics were used to compare the fit of

Table 5
Convergent validity correlations of scale scores with Model 6 factor scores.

Analysis	Factor				
	E	A	C	S	O
1	.923	.957	.929	.295	.833
2	.926	.763	.921	.584	.833
3	.920	.884	.829	.431	.894
4	.941	.879	.864	.099	.893
5	.864	.905	.935	.449	.925
6	.438	.552	.909	.826	.950

Table 6
Correlations between Big Five latent variables from application of Model 6. Factor determinacies are on the diagonal.

Analysis	1	2	3	4	5
1. Extraversion					
1	.939				
2	.913				
3	.945				
4	.887				
5	.925				
6	.857				
2. Agreeableness					
1	.195	.922			
2	.097	.854			
3	.290	.912			
4	.451	.792			
5	.289	.912			
6	−.381	.811			
3. Conscientiousness					
1	−.120	.197	.891		
2	−.021	.218	.907		
3	−.057	−.068	.894		
4	−.071	.101	.929		
5	−.221	.204	.917		
6	.092	.187	.920		
4. Stability					
1	−.410	−.253	−.257	.823	
2	.093	−.403	−.249	.856	
3	.098	.168	.033	.882	
4	−.100	−.147	.019	.968	
5	−.266	−.403	−.321	.862	
6	−.126	−.304	−.023	.902	
5. Openness					
1	.127	.319	−.070	.246	.878
2	−.043	−.053	.014	.147	.870
3	.231	.187	−.143	−.267	.871
4	.193	.246	−.190	.049	.804
5	.195	.085	−.014	.157	.912
6	−.179	−.217	−.296	−.015	.924

Model 6 with Big Five factor correlations estimated vs. the fit with the factors restricted to zero. The values of the chi-square statistics, each with $df = 10$, were 57.7, 30.3, 57.2, 68.7, 56.5, and 42.8 for the six analyses respectively ($p < .01$ for all). These results suggest that the Big Five dimensions as represented by the items in the IPIP and NEO questionnaires are not precisely orthogonal even when method variance is taken into account.

Although the chi-square values suggested a lack of precise orthogonality, the values of the chi-square difference statistics with $df = 10$ were small enough to warrant further examination of the practical consequences of restricting the correlations of the

Big Five factors to zero. Thus, a final examination of the issue of orthogonality was conducted in which factor scores from Model 6 where Big Five correlations were estimated were correlated with factor scores from Model 6 with Big Five factors restricted to be precisely orthogonal. Across the six analyses mean correlations between corresponding Big Five factors were .99, .97, .99, .97, and .99 for E, A, C, S, and O, and .99, .99, and .99 for M, Mp, and Mn respectively suggesting that factor scores computed from Model 6 with the restriction that the Big Five dimensions be orthogonal would be nearly equivalent to those computed estimating the Big Five correlations.

3.5. Impact of M

An example of the possible substantive importance of M and at the same time of the impact of method factors on correlations of Big Five variables with other variables was available from the data of sample 2. In this sample, the PANAS (Watson, Clar, & Tellegen, 1988) was administered to all participants under instructions to respond honestly. The PANAS instructions were to respond indicating how they generally felt, that is, how they felt on average. Recent research has raised the possibility that under instructions to respond honestly, the general method factor might be a measure of self-concept such as that represented by the two scores obtained from the PANAS – positive affectivity and negative affectivity (Biderman, Nguyen, & Cunningham, 2011; see also Loehlin & Martin, 2011). Biderman et al. (2011) found that M, estimated from Model 2 correlated positively with the Rosenberg Self Esteem scale and negatively with the Costello and Comrey (1967) Depression scale. These results suggested that M would correlate positively with measures of positive affectivity and at the same time correlate negatively with measures of negative affectivity. To investigate this possibility, we computed measures of positive affectivity and negative affectivity from the PANAS scale administered in sample two.

Since M has been found to influence all the Big Five items, if M is correlated with the PANAS variables, the influence of M on the Big Five items would result in the Big Five scales also being correlated with the PANAS variables. Table 7 presents the correlations of positive affect and negative affect with the Big Five scale scores. As can be seen, all Big Five scales correlated positively with positive affectivity and negatively with negatively affectivity ($p < .05$ for each).

In the middle panel of Table 7 are presented the correlations of both positive affectivity and negative affectivity with factor scores of all eight factors from the application of Model 6. The first important result in the table is that the general bias factor, M, correlates positively with positive affectivity and negatively with negative

Table 7
From sample 2, correlations of PANAS positive and negative affect scales with Big Five scale scores, with Big Five Model 6 factor scores, and with correlations of Big Five scale scores partialling out Model 6 M factor scores.

	Factor							
	E	A	C	S	O	M	Mp	Mn
Correlations with Big Five scale scores								
Positive affect	.335[c]	.229[b]	.228[b]	.323[c]	.346[c]			
Negative affect	−.235[b]	−.174[a]	−.183[b]	−.628[c]	−.187[b]			
Correlations with Model 6 Big Five factor scores								
Positive affect	.170[a]	−.059	.080	.156[a]	.115	.465[c]	−.025	.098
Negative affect	−.087	.115	−.048	−.327[c]	.004	−.361[c]	.062	−.496[c]
Partial correlations with Big Five scale scores partialling Model 6 M factor scores								
Positive affect	.179[a]	−.098	.110	.157[a]	.120			
Negative affect	−.103	.077	−.088	−.562[c]	.017			

[a] $p < .05$.
[b] $p < .01$.
[c] $p < .001$.

affectivity, replicating the finding of Biderman et al. (2011) and extending them to the PANAS measures of affectivity. Secondly, the correlations of factor scores from Model 6, from which the effect of method biases had been removed, are substantially different than the corresponding correlations of the Big Five scales. Specifically all the positive correlations were less positive and the negative correlations less negative, as would be expected if the effect of a variable with the pattern of correlations exhibited by M were removed. The changes were enough to alter significance conclusions for seven of the ten correlations.

The bottom panel of Table 7 illustrates an alternative way of removing the influence of method factors on correlations with external variables. In that panel are shown partial correlations of scale scores with both PANAS variables, partialling out M factor scores from application of Model 6. Again, the correlations were reduced in the expected directions, with only three correlations exceeding the significance threshold as above. The correlations of M with both positive affectivity and negative affectivity suggest that this "method" factor may represent substantive individual differences in the respondents. The changes in the conclusions concerning correlations of scale scores with the two PANAS variables emphasizes the fact that taking into account the influence of method factors can dramatically alter estimates of how strongly the Big Five dimensions are correlated with other variables.

4. Discussion

The objectives of the present study were to investigate the extent to which the data of two widely used Big Five personality trait questionnaires were affected by common method biases and to examine the nature and implications of such biases. Results suggest that the measurement model of the Big Five should take into account two types of method bias – one general bias factor influencing all items and a second type of bias factor influencing items worded either positively or negatively. The support for the existence of factors specific to item wording is in accordance with recent results presented by Marsh et al. (2010) for the RSE. The results suggest that a CFA of Big Five data using the IPIP or NEO measures may be misspecified if it does not include both a general method bias factor and factors associated with item wording.

The finding of method bias effects naturally leads to the question of the importance of those effects for theory and practice. The results of the test of hypothesis 3 suggest that the issue of the presence and effects of method biases has clear implications for theories of the relationships of the Big Five personality traits. Our results extend the work of Paglis and Williams (1996) and Williams and Brown (1994) on the effects of method factors on estimates of latent variable correlations. We found that estimates of the latent variable correlations were less positive than scale score correlations. We further found that when method bias was taken into account, factor scores computed under the assumption of completely orthogonal Big Five latent variables were nearly equivalent to factor scores from a model in which the latent variable correlations were estimated. That the correlations of the Big Five latent variables may be essentially, if not precisely, zero provides little support for the theories of higher order factors between the Big Five dimensions (DeYoung, Peterson, & Higgins, 2001; Digman, 1997; Musek, 2007). Many of the results previously presented in support of higher order personality factors have been based on data in which method bias factors were not estimated. In this study, in which the influence of method factors was taken into account, evidence for the positive correlations among the Big Five traits needed to be able to treat them as indicators of higher order factors was minimal, suggesting that the Big Five dimensions may be so nearly orthogonal that there could be no higher order factors.

The method factors estimated here, including the general method factor, M, may actually coexist with higher order factors (Chen et al., 2006). Fig. 2 presents a model in which both method factors and higher order factors are represented. On the left in Fig. 2 is Model 6 as applied here to the IPIP items. On the right is the higher order factor structure assumed by Digman (1997) and DeYoung et al. (2001). It seems that the model of Fig. 2 would be the model most appropriate to provide evidence for higher order factors. Fig. 2 shows that the method factors estimated in this study are orthogonal to the Big Five factors. Thus, they do not represent a conglomerate of the Big Five factors; they are influences on responses that are completely separate from the Big Five. This means that M, for example, is not the same factor as the general factor of personality (GFP) considered by Musek (2007) and represented on the right side of Fig. 2.

We point out that a model assuming no method factor, but estimating a general factor of personality – the right side of Fig. 2 – is a special case of Model 1, the worst fitting model applied here. Model 1 allowed the correlations between the Big Five factors to be estimated freely. Any model assuming higher order factors would require restrictions to that model, worsening the fit relative to the already poorly fitting Model 1.

The practical implications of the presence of method biases may also be important. In spite of the general recognition of the possibility of correlations being distorted due to the influence of a general bias, there have been few instances in which changes in correlations after removal of the influence of bias have been reported. In one of those, Biderman, Nguyen, and Sebren (2008) found that that the validity of conscientiousness as a predictor of test scores was increased substantially when factor scores from Model 2 were used in place of scale scores. The same result was found from unpublished analyses of data from sample 5 in the present study (Biderman et al., 2009) in which the validity of conscientiousness as a predictor of GPA measured by both the IPIP and the NEO-FFI questionnaires was larger when estimated by Model 2 factor scores than by scale scores. The impact of failure to account for method bias was also shown in the analyses reported above in which the correlations of PANAS variables with the Big Five dimensions were substantially different when factor scores from Model 6 were used rather than scale scores. These results all suggest that taking into account method biases can substantially change the estimated relationships of Big Five dimensions with other variables.

4.1. Common influence

The results of the chi-square difference tests supporting hypotheses 1 and 2 suggest that there is a common influence present in these samples, one that influences individuals' responses to the items in this personality inventory. In other words, within a group of persons sharing the same levels of personality characteristics, our results suggest that some individuals will report higher levels of all the traits, while others will report lower levels. It should be noted that these individual differences are not consistent with a form of simple acquiescence bias. If M represented simply individual differences in acquiescence, the tendency to agree with positively worded items would be canceled out by the tendency to agree (acquiesce) with negatively worded items, and the overall influence of M would be essentially zero. Instead of individual differences in agreement with items as written, M appears to represent individual differences in reporting of the extent to which the desirable aspect of the item – the actual item in the case of those that are positively worded, but its negation in the case of those that are negatively worded – represents the respondent.

One possible source of the observed common variation across all items is distortion that has been called socially desirable responding, dissimulation, or faking. As already mentioned, several
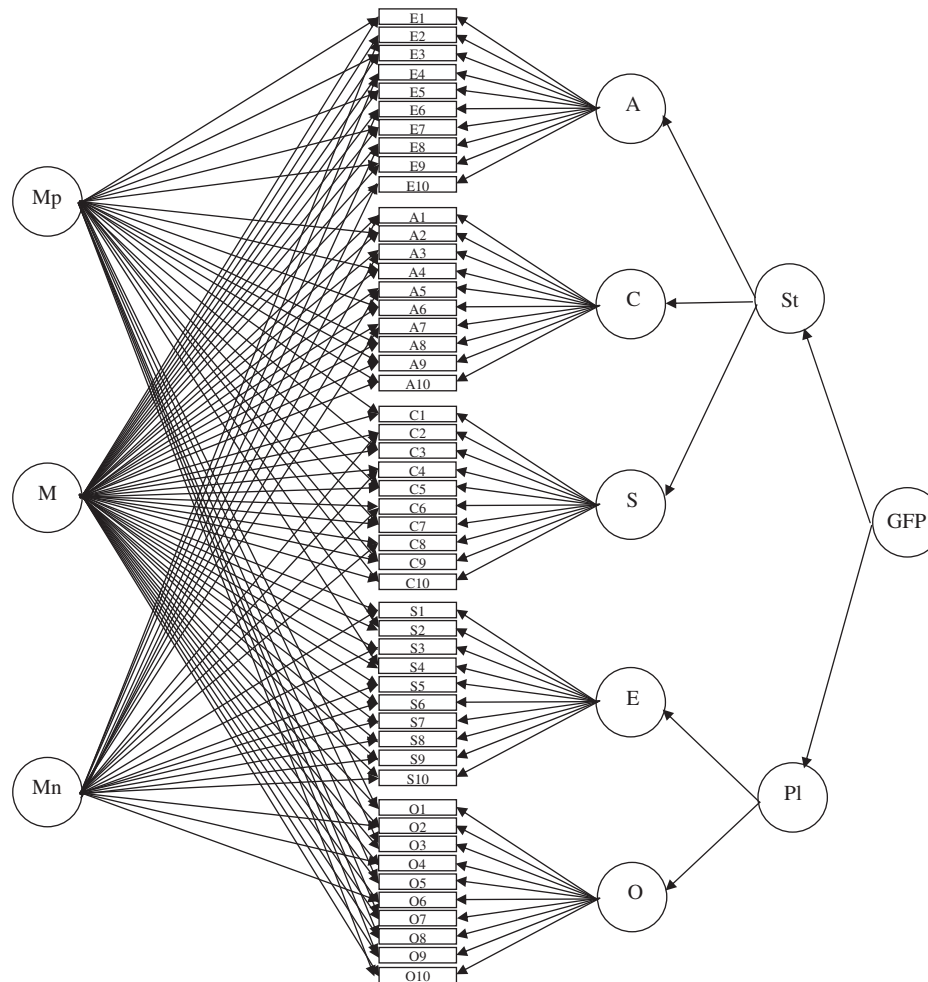
Fig. 2. The relationship of Model 6 applied to IPIP data and a model of the general factor of personality based on higher order factors. Residuals have been omitted for clarity.

investigators have found that when incentives or instructions to fake are present, common method variance analogous to that represented by M in this study may be due to differences in such responding (Biderman & Nguyen, 2004, 2009; Bäckström, 2007; Bäckström et al., 2009; Cellar et al., 1996; Klehe et al., 2011). Bäckström et al. (2009) found strong correlations between M and social desirability scales formed from the IPIP item pool. Biderman and Nguyen (2009) found strong correlations between difference scores computed by subtracting means in an honest response condition from means in a faking condition and M estimated from only the condition in which incentives to fake were present. This result indicated convergent validity of M with difference scores, often used to measure faking (McFarland & Ryan, 2000, 2006). More importantly for the present study, Biderman and Nguyen (2009) found positive correlations between M estimated from the honest condition and the incentive conditions of their study. This result indicated that respondents who faked the most when incentives were present were also likely to exhibit positive biases when instructed to respond honestly. Thus, even though respondents in all five samples included in this study were told to respond honestly, it is possible that there were individual differences in socially desirable responding such as those found by Bäckström et al. (2009) and Biderman and Nguyen (2009).

Another possibility is that under conditions to respond honestly, M may reflect personal affective states or self concept. This possibility is suggested by the correlations of M with the PANAS variables in the present study and by the similar pattern of corre-

lations of M with self esteem and depression scales found by Biderman et al. (2011). Thus it is possible that the biases found in the present study represent individual differences in a generalized affective disposition among respondents (Loehlin & Martin, 2011). This possibility may explain results of a concurrent validity study by Biderman, Nguyen, Mullins, and Luna (2008) in which the IPIP Big Five questionnaire was given to 764 incumbents in a financial services company. Supervisor ratings of employee performance on three dimensions – sales, customer service, and collections – were the criteria. When the criteria were regressed onto Big Five scale scores, no consistent predictor of performance in any criterion was found, and the multiple R was not significantly different from zero in any analysis. However when Big Five factor scores along with factor scores of M from Model 2 were treated as predictors, the multiple R was significant for each criterion. Moreover, the only individual predictor that was significantly related to all three criteria was M. This result is consistent with the suggestion that M represents self report of affective state and that those employees who exhibited the most positive affective states were rated best on all criteria by supervisors.

The possibility that M represents self report of affective state is not inconsistent with the possibility that M represents dissimulation in situations in which there are incentives to fake. When responding to a personality item with clear valence, for which a positive response indicates more of a desirable characteristic, biasing the response positively would be expected from a respondent wishing to appear more desirable in an applicant situation. But

biasing the response positively would also be expected from a respondent with a positive self concept while biasing it negatively would be expected from a respondent whose self concept was negative. Thus different root causes – the need to appear desirable in one instance and the tendency to report feelings about oneself in the other – may lead to similar individual differences in M.

4.2. Differences between Mp and Mn

The demonstration of the need for separate item-wording factors in this study of the Big Five and in the Marsh et al. (2010) study involving a different questionnaire raises the question of why there are two separate method effects. One possibility is that these differences could be due to the heightened cognitive drain associated with processing negatively vs. positively worded items (Cordery & Sevastos, 1993; Schmitt & Stults, 1985; Schriesheim & Hill, 1981). If this is the case, then the Mp and Mn effects observed here would not necessarily represent meaningful individual differences in personality but rather an effect of the measurement method that impacts the manner in which people respond to these types of items.

If the differences between bias represented by Mp and bias represented by Mn are due only to difficulty in processing the negation associated with negatively worded items, then it would be expected that persons lower in cognitive ability would exhibit larger differences between scores on Mp and Mn than respondents higher in cognitive ability. Some evidence regarding this cognitive drain hypothesis is available from samples 1 and 2 used here whose procedures included administration of the Wonderlic Personnel Test, a common measure of general cognitive ability (Wonderlic, 1999). For these two samples, the absolute difference, |Mp–Mn|, was computed from the factor scores of application of Model 7. These absolute differences were correlated with WPT scores for each dataset. The correlations were $-.184$ ($p < .05$) for sample 1 and $-.215$ ($p < .01$) for sample 2, with both negative correlations suggesting that the differences were smaller for persons with higher cognitive ability. Thus, these results provide some support for the hypothesis that the differences between Mp and Mn may be at least in part due to difficulty in processing the negatively worded items. The results are in agreement with the finding of Schmitt and Stults (1985) using simulated data who reported it only took 10% of participants to respond randomly for a method factor to emerge that would be related to negatively worded items.

An alternative explanation for these observed effects may be found in terms of an individual's underlying level of NA. It is conceivable that NA operates more directly on negatively worded items than on positively worded items. Such differences could lead to the need to represent the two item types by two different factors. Thus, the affectivity hypothesis might account for both the common method factor, M, and for Mp and Mn and the difference between them.

Both of the above possibilities focus on the Mn factor, failing to provide a reason for the existence of a bias associated only with positively worded items. At the present time, we have no convincing explanation for the existence of such a bias. It is possible that it is acquiescence affecting primarily positive worded items that does not affect items that are negatively worded.

The evidence for the existence of influences specific to negatively worded items and of different influences specific to positively worded items means that it may be important for designers of questionnaires to take note of the balance of positively and negatively worded items across dimensions. This is because the scale scores of dimensions indicated by mostly negative worded items (the IPIP Stability scale, for example, with eight such items) will reflect not only the trait but also be substantially influenced by Mn. Scale scores comprised of mostly positively worded items will be substantially influence by Mp. Such differences in method factor influence may confound differences between difference dimensions with a questionnaire – two dimensions with mostly positively worded items will be more highly correlated than two dimensions with different wording direction just due to differences in the influence of Mp and Mn. Difference in method factor influence may also confound correlations between similar dimensions across questionnaires measuring the same dimensions, such as the IPIP and NEO used here. Accounting for the differences between these biases will likely require specially constructed questionnaires with a careful balance of positively and negatively worded items for each dimension. Further examination of the differences using the existing questionnaires with different numbers of positively and negatively worded items per dimension is beyond the scope of the present research.

Although the term "method" is used to label the M, Mp, and Mn factors, the label refers to the fact that the factors are estimable only when a single method is used across several dimensions. The source of the bias itself is not the method but the respondent. These respondent influences are not apparent and estimable unless multiple personality dimensions are assessed using a single method and analyzed with items or parcels as indicators. Thus, we do not consider the influences reported in this study to be statistical artifacts. Ultimately they arise from characteristics of the respondents, and variance due to them is the result of individual differences in those personal characteristics. The source of method variance is individual differences in respondents, just as the source of extraversion variance, for example, is individual differences in respondents.

The case that M, Mp, and Mn are respondent characteristics is supported by the correlations of M with measures of affectivity and self concept and by the large convergent validities across the IPIP and NEO instruments in sample 5. These validities were of the same magnitude of the Big Five factors, whose origins are not in question. The convergent validities found in this study are in agreement with those found in the longitudinal study of Marsh et al. (2010). Clearly, further assessment employing longitudinal studies using Big Five data is in order.

4.3. Limitations and future directions

Because the orthogonality of the factors in the models presented in this study depends on the specific indicators, further research modeling other Big Five questionnaires with method factors is needed. The present study has provided initial evidence that bias is present in two widely used personality assessments. Whether other assessments are susceptible to similar bias is not known. For example, recent attempts to create Big Five scales from items taken from another widely used personality assessment, the CPI, have yielded scales whose correlations were quite close to zero (Soto & John, 2009) which would imply small effects of method variance. It would be interesting to verify this by modeling CPI items in a fashion similar to that done in this study.

The results of studies like the present one might be used to create a Big Five assessment that is free of the method effects. Taking note of differences in factor loadings of items on the method factor, it may be possible to select items that have high loadings on the Big Five traits while at the same time have small loadings on M, Mp, and Mn. In fact, Bäckström et al. (2009) were able to modify items in the IPIP 100-item Big Five questionnaire so that their loadings on a single method factor were considerably smaller than the loadings of the originally-worded items. Several iterations of such a selection process might yield a questionnaire for which the issue of method bias was considerably diminished or eliminated.

Those familiar with structural equation modeling may have noticed that the goodness-of-fit measures, particularly the CFI mea-

sures reported in Table 2, were poorer than is often reported in CFAs of questionnaires such as those used in this study. We believe that the poor fit of the models may be due in large part to the models' failure to account for idiosyncratic correlations between the individual items that served as indicators in the analyses. In traditional research involving correlations among scale or domain scores, these idiosyncrasies are buried within the scales. The analysis of items, however, bares the idiosyncrasies, and the goodness-of-fit measures, not included in traditional scale/domain score research, call attention to them. These poor goodness-of-fit statistics may raise the bar for the creation of new measures of personality and other traits, requiring investigators to pay closer attention to individual item correlations than has been done in the past. For example, Hall, Snell, and Foust (1999) called attention to the problem of failing to model secondary constructs in structural equation modeling.

There is much evidence that combining items into parcels results in models that exhibit better goodness-of-fit statistics than models of the same data using items as indicators (e.g., Kenny & McCoach, 2003; Lim & Ployhart, 2006; Thompson & Melancon, 1996). This means it is possible to model Big Five data so that goodness-of-fit measures such as CFI, RMSEA, and SRMR will exceed recommended thresholds as shown by Lim and Ployhart (2006). However, combining positively and negatively worded items within the same parcels or scales will likely obscure one of the key findings of the present research – the differences in bias between positively and negatively worded items. For that reason, unless parcels are created to reflect item wording, use of parcels or scales composed of mixtures of positively and negatively worded items will result in models that are misspecified to some extent.

Recently, Marsh et al. (2010) used Exploratory Structural Equation Modeling (ESEM) techniques and accounted for factor correlations in terms of correlated uniquenesses between items of the NEO-FFI. They found the Big Five factors to be essentially orthogonal after estimating 57 correlated uniquenesses that had been identified based on a priori considerations. Thus their approach to accounting for correlations between the Big Five factors represents an alternative to the method factors proposed here. We believe that there are strong arguments favoring the present models. These include the relationship of M to faking and to affective states. It is difficult to see how such relationships would be accounted for through correlated uniquenesses. The high convergent validity of the method factors across the two different questionnaires from sample 5 would also be hard to account for in terms of correlated uniquenesses since the questionnaires were based on different items. A comparison of method bias and correlated uniqueness explanations of factor correlations is beyond the scope the present research but clearly represents an avenue for future research.

The ubiquity of common method variance and the need to account for it by including method factors may also change the way research involving personality questionnaires is conducted. Even if an investigator wishes to examine relationships involving only one of the Big Five personality dimensions, that investigator may need to administer a complete Big Five questionnaire to be able to estimate method bias and remove its effect on the Big Five variable of interest as done in the partial correlations computed in the bottom panel of Table 7. In situations in which method bias is believed to contaminate both independent and dependent variables in a relationship, it is even conceivable that the Big Five questionnaire could be administered solely for the purpose of obtaining method bias factor scores, for example, so that they could be partialled from analyses of relationships in which direct estimation of method effects would not be possible.

5. Conclusions

Common method variance 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. Now may be the time to begin to understand it and to investigate the extent to which it affects other aspects of behavior. The payoff may be the understanding of characteristics of personality that have been hidden in personality questionnaires all along.

References

Bäckström, M. (2007). Higher-order factors in a five-factor personality inventory and its relation to social desirability. *European Journal of Psychological Assessment, 23*, 63–70.

Bäckström, M., Björklund, F., & Larsson, M. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality, 43*, 335–344.

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

Biderman, M. D., & Nguyen, N. T. (2009). Measuring faking propensity. In *Paper presented at the 24rd annual conference of the society for industrial and organizational psychology*, New Orleans, LA.

Biderman, M. D., Nguyen, N. T., Mullins, B., & Luna, J. (2008). A method factor predictor of performance ratings. In *Paper presented at the 23rd annual conference of the society for industrial & organizational psychology*.

Biderman, M. D., Nguyen, N. T., & Cunningham, C. L. (2009). Common method variance in NEO-FFI and IPIP personality measurement. In *Paper presented at the 24th annual conference of the society for industrial & organizational psychology*.

Biderman, M. D., Nguyen, N. T., & Cunningham, C. L. (2011). A method factor measure of self-concept. In *Paper accepted for presentation at the 26th annual conference of the society for industrial & organizational psychology*.

Biderman, M. D., Nguyen, N. T., & Sebren, J. (2008). Time-on-task mediates the conscientiousness–performance relationship. *Personality and Individual Differences, 44*, 887–897.

Bollen, K. A., & Paxton, P. (1998). Detection and determinants of bias in subjective measures. *American Sociological Review, 63*, 465–478.

Cellar, D. F., Miller, M. L., Doverspike, D. D., & Klawsky, J. D. (1996). Comparison of factor structures and criterion-related validity coefficients for two measures of personality based on the five factor model. *Journal of Applied Psychology, 81*, 694–704.

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

Cordery, J. L., & Sevastos, P. P. (1993). Responses to the original and revised Job Diagnostic Survey: Is education a factor in responses to negatively worded items? *Journal of Applied Psychology, 78*, 141–143.

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

Costello, C. G., & Comrey, A. L. (1967). Scales for measuring depression and anxiety. *The Journal of Psychology, 66*, 303–313.

Cote, J. A., & Buckley, R. (1987). Estimating trait, method, and error variance. Generalizing across 70 construct validation studies. *Journal of Marketing Research, 24*, 315–318.

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 fakeability of personality questionnaires: Faking for specific jobs. Unpublished master's thesis. University of Tennessee at Chattanooga, Chattanooga, TN.

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage.

DeYoung, 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.

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

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

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., & Somer, O. (2000). The hierarchical structure of common Turkish person-descriptive adjectives. *European Journal of Personality, 14*, 497–531.

Grice, J. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*, 430–450.

Hackman, J. R., & Oldham, G. R. (1975). Development of the job diagnostic survey. *Journal of Applied Psychology, 60*, 159–170.

Hall, R. J., Snell, A. F., & Foust, M. S. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods, 2*, 233–256.

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.

Idaszak, J. R., & Drasgow, F. (1987). A revision of the Job Diagnostic Survey: Elimination of a measurement artifact. *Journal of Applied Psychology, 72*, 69–74.

Kenny, D. A., & McCoach, D. B. (2003). Effect of number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling, 10*, 333–351.

Klehe, U. -C. et al. (2011). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. Manuscript under review.

Lim, B., & 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.

Loehlin, J. C., & Martin, N. G. (2011). The general factor of personality: Questions and elaborations. *Journal of Research in Personality, 45*, 44–49.

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 artifacts? *Journal of Personality and Social Psychology, 70*, 810–819.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., et al. (2010). A new look at the Big Five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22*, 471–491.

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment, 22*, 366–381.

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.

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.

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

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.

Mount, M. K., Barrick, M. R., Scullen, S. M., & Rounds, J. (2005). Higher order dimensions of the big five personality traits and the big six vocational interest types. *Personnel Psychology, 58*, 447–478.

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, B.O. (1998–2004). Mplus technical appendices. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (1998–2006). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.

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

Paglis, L. L., & William, L. J. (1996). Common method variance: When does it bias OB research results? In *Paper presented at the meeting of the Society for Industrial and Organizational Psychology*, San Diego, CA (April).

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.

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

Rushton, J. P., Bons, T. A., & Hur, Y. (2008). The genetics and evolution of the general factor of personality. *Journal of Research in Personality, 42*, 1173–1185.

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. (2003). The structure of personality attributes. In M. R. Barrick & A. M. Ryan (Eds.), *Personality and work: Reconsidering the role of personality in organizations* (pp. 1–29). San Francisco, CA: Jossey-Bass.

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., & 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.

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.

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.

Soto, C. J., & John, O. P. (2009). Using the California Psychological Inventory to assess the Big Five personality domains: A hierarchical approach. *Journal of Research in Personality, 43*, 25–38.

Spector, P. E. (1988). Development of the work locus of control scale. *Journal of Occupational Psychology, 61*, 335–340.

Spector, P. E. (1998). *Summated rating scale construction*. Thousand Oaks, CA: Sage.

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

Tomas, J. M., Hontangas, P. M., & Oliver, A. (2000). Linear confirmatory factor models to evaluate multitrait–multimethod matrices: The effects of number of indicators and correlation among methods. *Multivariate Behavioral Research, 35*, 469–499.

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

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

Watson, D., Clar, L., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063–1070.

Williams, L. J., & Brown, B. K. (1994). Method variance in organizational behavior and human resources research: effects on correlations, path coefficients, and hypothesis testing. *Organizational Behavior and Human Decision Processes, 57*, 185–209.

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.

Wonderlic Inc. (1999). *Wonderlic's personnel test manual and scoring guide*. Chicago, IL: Author.

Ziegler, M., & Buehner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement, 69*, 548–565.