Implications of Item Keying and Item Valence for the Investigation of Construct Dimensionality

Chester Chun Seng Kam

University of Macau

John P. Meyer

The University of Western Ontario

Factor analysis and nomological network analysis are commonly used as complementary procedures in the investigation of the dimensionality of constructs (e.g., self-esteem, job satisfaction). Although it has been demonstrated that factor analyses are often biased toward a two-dimensional solution for measures including regular- and reverse-keyed items, less attention has been paid to the implications for nomological network analyses. We propose, and demonstrate empirically in two studies, that item keying is confounded with item valence (i.e., favorability of item content), and that item valence can bias the results of both factor analysis and nomological network analysis toward a two-dimensional interpretation. We also demonstrate that the valence effect is related to, but distinguishable from, social desirability response bias. We caution that the practice of excluding reverse-keyed items to achieve unidimensionality can lead to distortion in correlations among constructs, and we offer alternative remedies to the valence problem.

The accurate measurement of psychological constructs is essential for research and the advancement of knowledge. A key to effective measurement is a clear conceptualization of the construct. However, we often rely on measurement research to help formulate our ideas about the true nature of a construct and its association with other constructs. Thus, theory and measurement research are inextricably intertwined (Cronbach & Meehl, 1955). This link is clearly reflected in ongoing debates regarding the dimensionality of several important constructs across multiple areas of psychology, including presence versus absence of anxiety in clinical psychology (Vagg, Spielberger, & O’Hearn, 1980), positive versus negative self-esteem in personality psychology (Carmines & Zeller, 1979; Horan, DiStefano, & Motl, 2003; Quilty, Oakman, & Risko, 2006; Tomás & Oliver, 1999), dominance versus egalitarianism in personality and social psychology (Ho et al., 2012; Pula, McPherson, & Parks, 2012), optimism versus pessimism in personality and clinical psychology (e.g., Marshall, Wortman, Kusulas, Hervig, & Vickers, 1992), marital satisfaction versus dissatisfaction in family psychology (Fincham & Linfield, 1997), belief in the just world versus belief in the unjust world in social psychology (Rubin & Peplau, 1975), and job satisfaction versus dissatisfaction in industrial-organizational psychology (e.g., Credé, Oleksandr, Bagraim, & Sully, 2009). To illustrate, consider the case of positive self-esteem versus negative self-esteem. Some scholars argue that self-esteem is a unidimensional construct, whereas others believe that positive versus negative self-esteem are distinguishable, albeit related, constructs (see Marsh, 1996, for a discussion of this debate). These diverging beliefs can have important implications for research pertaining to the development and consequences of self-esteem as well as for clinical practices.

Although there has been extensive work demonstrating that factor analysis often produces a “keying factor” when applied to measures including both positively and negatively worded items, many researchers continue to rely heavily on basic factor analyses to determine the dimensionality of a construct. They often follow up on these analyses by conducting nomological network analyses (e.g., Herzberg, Gläsermer, & Hoyer, 2006; Marshall et al., 1992; Rauch, Schweizer, & Moosbrugger, 2007) to demonstrate that scores reflecting the regular- and reverse-keyed items relate differently to external scales and therefore reflect distinct constructs. In
contrast, other researchers consider the bi-dimensionality of a measure to be an artifact of item-keying and recommend the exclusion of reverse-keyed items (e.g., Lindwall et al., 2012; Magazine, Williams, & Williams, 1996; Schriesheim & Eisenbach, 1995; Schriesheim, Eisenbach, & Hill, 1991; van Sonderen, Sanderman, & Coyne, 2013). We believe that these decisions are being made without a clear understanding of why item-keying has the effects it does.

Our objectives in this research are fourfold. First, we seek to clarify the concept of item keying by introducing the notion of item valence (favorability) as distinct from item-keying direction (regular and reversed). This distinction is important because, if the emergence of two factors is solely due to item-keying, it can be taken as evidence that the construct is bi-dimensional (i.e., the regular- and reverse-keyed items reflect distinguishable constructs). However, if the emergence of distinct factors reflects item valence, subscales created from regular and reverse-keyed items will confound content and valence and therefore no definitive conclusions can be drawn with regard to construct dimensionality.

Our second objective is to investigate the implications of the potential confounding of content and valence for the results of nomological network analyses often conducted to complement factor analytic evidence of bi-dimensionality. In nomological network analyses, scores reflecting the two dimensions observed in a factor analysis are correlated with other constructs presumed to be the antecedents, correlates, or consequences within a nomological network (e.g., Créde et al., 2009). Demonstration of a different pattern of correlations is taken as further evidence that the constructs are distinct. In this manner, researchers often treat factor analysis and nomological network analysis as independent sources of information pertaining to construct dimensionality. We provide evidence to illustrate that both analyses are subject to the influence of item valence.

Our third objective addresses the implications of the recent recommendation that researchers use measures with only regularly keyed items (e.g., Lindwall et al., 2012; Magazine et al., 1996; Schriesheim et al., 1991; Schriesheim & Eisenbach, 1995; van Sonderen et al., 2013). This recommendation is based to a large extent on the evidence of the effects of item-keying on dimensionality, but also on the desire to reduce survey length or increase the number of constructs measured in a study. We demonstrate that the decision to include or exclude reverse-keyed items can influence the correlation between measures in nomological network analyses due, in part, to the effect of item valence.

Finally, we investigate one potential cause of the item valence effect—social desirability response bias. We argue that the implications of item valence, or the favorability of item content, for scores on a measure will be influenced by respondents’ tendency to present themselves in a positive light (Paulhus, 1991). Social desirability response bias may therefore explain at least some of the valence effect.

## Construct Dimensionality Debate

### Item Valence and Item-Keying Direction

Although there may be exceptions (e.g., monochronicity vs. polychronicity; agency vs. communion; individualism vs. collectivism), it is often the case that the attribute identified at one end of a bipolar construct is more favorable than the attribute at the other end. For example, positive self-esteem [measured by items such as “I feel that I’m a person of worth” in Rosenberg (1965) Self-Esteem Scale (RSES)] is more favorable than negative self-esteem (measured by items such as “At times I think I am no good at all”). We use the term valence to refer to this characteristic of an attribute and describe favorable attributes as positively valenced and unfavorable attributes as negatively valenced. This valence reflects the subjective evaluation that has long been recognized as being inherent in the way we describe people, things, or events (e.g., Beauvois & Dubois, 2000; Peabody, 1967; Osgood et al., 1957; Peterson, 1965; Rosenberg & Olshan, 1970; Wiggins, 1973).

It is important to note that the evaluative aspect (i.e., valence) of items in a measure is distinct from their direction of keying. For example, when emotional stability is used as the label for one of the Big Five personality traits, the reverse-keyed items measure neuroticism—a negatively valenced construct. However, if the construct is labeled neuroticism, then the reverse-keyed items measure emotional stability—a positively valenced construct. This distinction between keying direction and valence has been blurred (or confused) in the past by authors who refer to regular-keyed items as “positively keyed” and reverse-keyed items as “negatively keyed.” More often than not, the negatively keyed items being described do indeed have negative valence (i.e., we tend to put positive labels on our constructs). However, this is not always the case, and therefore reverse-keyed items can have a positive valence when the construct under investigation is negatively valenced (e.g., neuroticism; burnout, counterproductive work behavior). Most of the previous studies attributed the emergence of a method factor simply to the inclusion of reverse-keyed items, but as we have already argued, item-keying and item valence are naturally confounded with each other (Peabody, 1967).

### How Can Valence Affect Both Factor Analysis and Nomological Network Analysis?

As mentioned, it is often assumed by researchers that nomological network analyses and factor analyses provide independent sources of information concerning construct dimensionality. To the contrary, we propose that both analyses can be influenced by item valence and lead to potentially erroneous conclusions regarding dimensionality. That is, construct scores can reflect both the content and valence of the items. In factor analysis, the existence of a valence component can lead to identification of two dimensions when...
both positively and negatively valenced items are included. In nomological network analyses, when the positively and negatively valenced items are combined separately to create subscales, correlations observed between constructs will reflect covariance in both content and valence. Consequently, interpretation of the correlations based on content can be biased by the valence component (i.e., correlations will be elevated when valence is similar and reduced when valence is dissimilar). For example, if factor analytic results lead to the creation of separate job satisfaction and job dissatisfaction scores, the latter might correlate more strongly than the former with a measure of counterproductive work behavior, not (only) because dissatisfaction is a better predictor of such behavior than is satisfaction, but because the measures of job dissatisfaction and counterproductive work behavior are both negatively valenced. Thus, when the valence effect leads to emergence of two factors in factor analysis, and this evidence is used to justify creation of subscales for use in nomological network analysis, the same valence effect can lead to further misinterpretation of the findings in support of construct bi-dimensionality (Credé et al., 2009).

To investigate the role of item valence on both factor analysis and nomological network analysis, the current study employed a multitrait-multimethod (MTMM) framework, specifically relying on the standard correlated trait-correlated method confirmatory factor analysis (CTCM CFA) model (Jöreskog, 1971; Kenny, 1976; Widaman, 1985). Some researchers regard the CTCM model as a faithful representation of the original MTMM conceptualization (Lance, Noble, & Scullen, 2002). Consider that we have a total $p$ observed (or manifested) variables, $t$ traits, and $m$ methods. The baseline model for participant $k$ is:

$$
y_k = \Lambda_T \eta_{Tk} + \epsilon_k, \quad (1)
$$

where $y_k$ is a $p \times 1$ vector for the scores of $p$ mean-centered observed variables from participant $k$, $\Lambda_T$ is a $p \times t$ matrix representing the factor loading of observed variables on $t$ traits, $\eta_{Tk}$ is a $t \times 1$ vector for $t$ trait factor scores from participant $k$, $\epsilon_k$ is a $p \times 1$ vector representing unique factors of $p$ observed variables. The baseline model, however, does not account for the method effects of positively and negatively valenced items. The CTCM model involving method factor scores for participant $k$ is as follows:

$$
y_k = \Lambda_T \eta_{Tk} + \Lambda_M \eta_{Mk} + \epsilon_k, \quad (2)
$$

where $\Lambda_M$ is a $p \times m$ matrix representing the factor loading of $p$ observed variables on $m$ methods, $\eta_{Mk}$ is a $m \times 1$ vector for $m$ method factor scores from participant $k$. With the assumption of independence among traits, methods, uniqueness and the assumption of independence among unique factors, Equations (1) and (2) have the following covariance structure respectively:

- Common Factor Model: $\Sigma = \Lambda_T \Phi_T \Lambda_T' + \Theta, \quad (3)$
- CTCM: $\Sigma = \Lambda_T \Phi_T \Lambda_T' + \Lambda_M \Phi_M \Lambda_M' + \Theta, \quad (4)$

where $\Sigma$ is a $p \times p$ covariance matrix for $p$ observed variables, $\Phi_T$ is a $t \times t$ covariance matrix of $t$ traits, $\Phi_M$ is a $m \times m$ covariance matrix of $m$ methods, and $\Theta$ is a $p \times p$ diagonal matrix for unique variances.

In the current study, there are two methods ($m$), namely the use of positively valenced items ($j$) and the use of negatively valenced items ($j'$) to measure the same trait $i$. The total number of traits ($t$) and the total number of observed/manifested variables ($p$), however, vary across our two empirical samples. For model identification, we will assume that all trait and method factor variances are standardized (i.e., set as 1s). An item valence effect is evident if Equation (4) fits the data better than Equation (3), using the usual chi-square difference statistic as the comparison indicator.

From the Equation (4), it is apparent that the correlation between two specific observed variables (one variable $y_{ij}$ measuring trait $i$ with method $j$ and another variable $y_{ij'}$ measuring trait $i'$ with method $j'$) is:

$$r(y_{ij}, y_{ij'}) = \lambda_{Tij} \lambda_{Ti'j'} \psi_{Ti'Ti'} + \lambda_{Mij} \lambda_{Mij'} \psi_{Mij'Mij'}, \quad (5)$$

where $\lambda_{Tij}$ and $\lambda_{Ti'j'}$ refer to the trait factor loadings for the two observed variables ($y_{ij}$ and $y_{ij'}$), $\lambda_{Mij}$ and $\lambda_{Mij'}$ refer to the method factor loadings for the two observed variables. $\psi_{Ti'Ti'}$ refers to the correlation between two different traits (trait $i$ and trait $i'$), and $\psi_{Mij'Mij'}$ refers to the correlation between two different methods (method $j$ and method $j'$). Equation (5) refers to the heterotrait-heteromethod (HTHM) correlation in MTMM.

In the case of monotrait-heteromethod (HTHM) correlation, Equation (5) is reduced to:

$$r(y_{ij}, y_{ij'}) = \lambda_{Tij} \lambda_{Ti'j'} + \lambda_{Mij} \lambda_{Mij'} \psi_{Mij'Mij'}, \quad (6)$$

Equation (6) explains why positively and negatively valenced items (heteromethod) load on two separate factors in exploratory factor analysis (EFA). In MTHM correlations, the two methods ($M_i$ and $M_j$) are usually not perfectly correlated (i.e., $\rho_{M_i'M_j'} < 1$). In contrast, when a researcher uses only items with the same valence to measure the same construct (i.e., the case of monotrait-monomethod), the method correlation ($\rho_{M_i'M_j'}$) in Equation (6) will be reduced to 1, resulting in a stronger correlation than the MTHM correlation. Weaker MTHM correlations (as in situations where oppositely valenced items are used to measure the same construct) bias EFA results in favor of bidimensionality.

In the case of heterotrait-monomethod (HTMM) correlation, Equation (5) is reduced to:

$$r(y_{ij}, y_{ij'}) = \lambda_{Tij} \lambda_{Ti'j'} \psi_{Ti'Ti'} + \lambda_{Mij} \lambda_{Mij'} \psi_{Mij'Mij'}, \quad (7)$$

Equations (5) and (7) explain why item and construct valence can bias construct correlations. The difference between Equation (5) and Equation (7) is the absence of the inter-method correlation term ($\psi_{M_i'M_j'}$) in Equation (7).
variables measure distinct constructs (heterotrait). However, in Equation (5) the observed variables represent heterogeneous measurement methods (HTHM; such as positively and negatively valenced items), while in Equation (7), the observed variables represent the same measurement method (HTMM; such as positively valenced items only). Comparison of the two equations reveals that the HTMM correlation in Equation (7) will be stronger than the HTHM correlation in Equation (5) when the correlation between two methods ($M_i$ and $M_j$) in Equation (5) are not perfectly correlated (i.e., $|\phi_{M_iM_j}| < 1$). For this reason, observed variables that load on the same method factor will have a stronger correlation than observed variables that load on two separate method factors. Of course, this result is based on observed variable correlations but it can be easily extended to construct correlations, because constructs are estimated by all of their corresponding observed variables. When two constructs load on the same method factor (e.g., positively valence factor), they should have stronger correlation with each other; when two constructs load on different method factors (one on positively valence factor and another on negatively valence factor), they should have weaker correlation with each other.

THE PRESENT STUDY

Our objective in the present research is to demonstrate the potential impact of item and construct valence in factor analyses and nomological network analyses, using the measurement of self-esteem and extraversion as an example. It is not our intention to resolve the debate about the dimensionality of these constructs, but rather to investigate the potential implications of valence as it pertains to measure development and substantive research in general. Some readers may think that the item valence effect can be entirely explained by social desirability response bias, i.e., the tendency to respond in a way that creates a more favorable impression (Paulhus, 1991). Indeed item valence and social desirability response styles may be related to each other, and our final research question was whether social desirability response bias can fully explain the effect of item valence. We expect that social desirability response style will correlate positively with the positive valence factor and negatively with the negative valence factor. Nonetheless, there may be other item response processes that contribute to the item valence effect. Therefore, our investigation of the role played by social desirability serves only as an initial exploration of the mechanisms involved in this effect.

METHOD

Participants

Introductory psychology students at a large Canadian university were assigned to one of two samples for an online survey. Sample 1 consisted of 1094 students (760 female and two unidentified; mean age = 18.45) and Sample 2 consisted of 1254 students (873 female and one unidentified; mean age = 18.38). One participant from Sample 1 and two participants from Sample 2 did not fill out any of the measures, and thus were excluded from the analyses.

Measures

For purposes of this investigation, we used data from measures obtained in the two different mass testing sessions (see below). All except one measure (Belief in Zero-Sum Resources) included equal numbers of positively and negatively valenced items. This allowed us to compute full-scale scores as well as positively and negatively valenced scale scores for each substantive construct. The reliabilities of the scales are good (see Table S1 in Supplemental Data).

Personality

The Big Five personality factors (conscientiousness, extraversion, agreeableness, openness to experience, and neuroticism) were measured with scales (NEO domain) taken from the international personality item pool (IPIP; Goldberg et al., 2006). Each factor consists of 10 items with a 5-point Likert-type scale from 1 (strongly disagree) to 5 (strongly agree).

Self-Esteem

The Rosenberg (1965) Self-Esteem Scale (RSES) consisted of 10 items, and measured respondents’ global evaluation of self-worth. Each item was measured with a 4-point scale from 1 (strongly disagree) to 4 (strongly agree). A sample item is “On the whole, I am satisfied with myself.”

Social Dominance Orientation (SDO)

The SDO scale measures respondents’ preference for inequality and hierarchical differentiation in a social context. SDO was measured by 16 items, developed by Sidanius and Pratto (2001), with a 7-point scale from 1 (strongly disagree) to 7 (strongly agree). A sample item is “To get ahead in life, it is sometimes necessary to step on other groups.”

Belief in a Dangerous World (BDW)

The BDW Scale (Altemeyer, 1988), included in Sample 1 only, was composed of 12 items, each of which was measured with a 9-point scale from −4 (very strongly disagree) to +4 (very strongly agree). BDW measured respondents’ belief that the world is a dangerous and threatening place. A sample item is “It seems that every year there are fewer and fewer truly respectable people, and more and more persons with no morals at all who threaten everyone else.”
Belief in Zero-Sum Resources (BZSR)

This 6-item BZSR Scale, included in Sample 1 only, is a revised and shortened version of the original BZSR measure by Esses, Jackson, and Armstrong (1998). BZSR measures one’s beliefs that immigrants are competing with Canadians for valuable resources in society. A sample item is “Money spent on social services for immigrants means less money for services for Canadians already living here.” The current measure contains one reverse-keyed item. Each item was measured with a 7-point scale from 1 (strongly disagree) to 7 (strongly agree).

Balanced Inventory of Desirable Responding (BIDR)

The BIDR scale (Paulhus, 1991) was included in Sample 2 to measure social desirability responding. It consisted of 38 items measured with a 5-point Likert-type scale from 1 (strongly disagree) to 5 (strongly agree) in our study. Half of the items are reverse-keyed. The original scale consists of two subscales, namely impression management and self-deception. Researchers (e.g., Li & Bagger, 2007; Paulhus, 1991) generally conceptualized impression management as intentional distortion of a self-image whereas self-deception as an unintentional propensity to exaggerate positive attributes. We obtained scores for impression management and self-deception by averaging participants’ ratings on relevant items (see Kam, 2013).

RESULTS AND DISCUSSION

All of the structural equation modeling (SEM) analyses were conducted with the program Mplus 6.12 (Muthén & Muthén, 1998–2010) with the maximum likelihood robust estimator (MLR). The MLR allows some deviation from the multivariate normality assumption. All the model comparisons were conducted with Satorra-Bentler (2001, 2010) scaled difference chi-square tests. Missing data were treated with the Full Information Maximum Likelihood method (FIML; Enders, 2001; Graham, 2009).

Basic Confirmatory Factor Analysis (CFA)

To illustrate the effect of item valence on construct dimensionality, we first tested the dimensionality of the constructs in the current study using basic confirmatory factor analyses. We compared two models, a one-factor model and a two-factor model. In the one-factor model, all item scores loaded on a common factor. In the two-factor model, positively valenced items and negatively valenced items loaded on two separate factors. Individual items were used in indicators in these basic CFAs. Sample 1 results are shown here (see Table S2 in Supplemental Data) because this sample included data for all the constructs used in the current study. Results from Sample 2 are similar and thus are not shown. A two-factor solution fits better than a one-factor solution for six of the eight constructs (all ps < .001). These six constructs are self-esteem, extraversion, conscientiousness, agreeableness, SDO, and BDW. A two-factor solution is not better than a one-factor solution for neuroticism (Δχ² = 0.18, Δdf = 1, p = .67; r_factors = −.99) and openness (Δχ² = 0.06, Δdf = 1, p = .80; r_factors = −1.00). Based on the scaled difference chi-square difference test and fit indices (TLI and CFI), it should also be noted that the two-factor solution fit substantially better for agreeableness (Δχ² = 185.83, TLI = .92 vs..79, CFI = .94 vs..84) and SDO (Δχ² = 240.57, TLI = .90 vs..77, CFI = .92 vs..80) than for other constructs. Overall, there was a tendency for items with opposing valence to load on two separate factors. This tendency was found even for constructs (e.g., agreeableness, BDW) that have always been treated as unidimensional theoretically and empirically.

Multitrait-Multimethod Confirmatory Factor Analysis

As mentioned, we used the correlated trait-correlated method (CTCM) MTMM model to investigate the impact of item valence. We compared a baseline model to a method-factor model. In the baseline model (Mbaseline), item indicators load on the intended construct factors only (see Figure 1). In the method-factor model (Mvalence), we included two valence (method) factors and allowed item indicators to load on both the intended construct factor and the corresponding valence factor (positive or negative). The two valence factors are also allowed to correlate with each other. If Mvalence fits the data better than Mbaseline, it supports the notion that shared valence contributes to the correlations among the items.

All construct factors and method factors were set to have a variance of unity, and all factor loadings were freely estimated. Before the analyses, all of the scale items were parcelled because MTMM CFA solutions often do not converge satisfactorily when too many item indicators are included in the model (Bentler & Chou, 1987; Yuan, Bentler, & Kano, 1997). Each parcel comprises 2–3 items. For constructs measured by five or fewer items for each valence (e.g., positive self-esteem, negative conscientiousness), we could only create two parcels (except positive BZSR, which

---

1Two items relevant to sex-relevant behaviors and love-related cognition were removed from the data collection due to concerns from the ethics board.

2The dimensionality of BZSR in Sample 1 could not be tested with CFA because it contains only one positively valence item.

3For each construct, positively valenced items and negatively valenced items are parcelled separately.
only consists of one measurement item). For constructs measured by more than five items for each valence (e.g., negative BDW, positive SDO), we created three parcels. In addition, each of the constructs, with the exception of BZSR which has only one positively valenced item, has the same number of positive parcels and negative parcels. Recently, Sterba and MacCallum (2010) demonstrated that methods of parcel-allocation can cause variability in CFA model fit. Therefore, they recommended that researchers report and control for parcel-allocation variability. In the initial step, a SAS macro that was provided by Sterba automatically generated a number of data sets (100 sets in the present study) with random item-to-parcel allocations. In the second step, Mplus examined the averaged model estimates based on these data sets. Model fit of these averaged fit indices across the 100 data sets are shown in Table 1. In addition, variability of fit indices was found to be small in both Samples 1 and 2, as the standard deviations of these fit indices represent less than 2% of their corresponding fit indices’ values (see Supplementary Table S3 for the reporting of these standard deviations). This

![Figure 1](image-url)

**FIGURE 1** Models tested in the current study. PL = parcel for positively valenced items; NL = parcel for negatively valenced items. Belief in Zero-Sum Resource Scale (BZSR) and Belief in a Dangerous World Scale (BDW) were only in Sample 1 and not in Sample 2.

| Sample 1                      | Sample 2                      |
|-------------------------------|-------------------------------|
| Baseline model                | Baseline model                |
| % Successful convergence     | % Successful convergence     |
| Across Data Sets              | Across Data Sets              |
| χ²                            | χ²                            |
| df                            | df                            |
| p                             | p                             |
| TLI                           | TLI                           |
| CFI                           | CFI                           |
| RMSEA                         | RMSEA                         |
| SRMR                          | SRMR                          |
| AIC                           | AIC                           |
| SABIC                         | SABIC                         |

| Sample 1                      | Sample 2                      |
|-------------------------------|-------------------------------|
| Baseline model                | Baseline model                |
| % Successful convergence     | % Successful convergence     |
| Across Data Sets              | Across Data Sets              |
| χ²                            | χ²                            |
| df                            | df                            |
| p                             | p                             |
| TLI                           | TLI                           |
| CFI                           | CFI                           |
| RMSEA                         | RMSEA                         |
| SRMR                          | SRMR                          |
| AIC                           | AIC                           |
| SABIC                         | SABIC                         |

**TABLE 1** Averaged Model Fit Estimate in Both Sample 1 and 2

Note. TLI: Tucker Lewis Index; CFI: Comparative Fit Index; RMSEA: Root Mean Square Error of Approximation; SRMR: Standardized Root Mean Square Residual; AIC: Akaike Information Criterion; SABIC: Sample-size Adjusted Bayesian Information Criterion.

Averaged fit indices across 100 parcel-level data sets from a random item-to-parcel allocation distribution. Maximum likelihood estimator with robust standard errors is used in these test results.

M₂ valence = Model with two valence (method) factors.
result was not surprising as Sterba and MacCallum (2010) have shown that a large sample size is resistant to parcel-allocation variability.

$M_{\text{baseline}}$ was nested within $M_{2\ \text{valence}}$. The fit of $M_{2\ \text{valence}}$ was a significant improvement over that for $M_{\text{baseline}}$ based on scaled chi-square difference test in Sample 1 ($\Delta \chi^2 = 1130.16$, $\Delta df = 40$, $p < .001$) and Sample 2 ($\Delta \chi^2 = 1275.48$, $\Delta df = 31$, $p < .001$), indicating the existence of the two valence factors (see Table 1 for fit indices). Therefore, across two independent samples our results are consistent with the hypothesis that correlations among the scale items reflect both their content and their valence. Thus, our findings suggest that the inclusion of positively and negatively valenced items in the measure of a construct has the potential to lead to the extraction of two factors even under conditions when the items may reflect common content.

The correlation between the positive valence factor and the negative valence factor is quite negative in Sample 1 ($r = -.50$, $SD = .05$ across randomly parceled data sets). The same correlation is slightly negative in Sample 2 ($r = -.10$, $SD = .10$ across randomly parceled data sets). Although the correlations differ between two samples, they were consistently far from a perfect negative correlation. This result suggests that participants do not have a consistent response style between positively valenced items and negatively valenced items (Geiser, Eid, & Nussbeck, 2008). Thus we did not assume only one latent factor to account for the method variance due to valence in the current data (Maydeu-Olivares & Coffman, 2006). The factor loadings of the MTMM model in the two samples are shown in Supplementary Table S4.

Nomological Network Analyses and Correlational Analyses

To determine how construct valence affects the correlations of measures with external antecedents, correlates or consequences, we compared how positive and negative self-esteem, and extraversion and introversion correlated with the other constructs measured in Samples 1 and 2. Construct scores differ in their reliability (Meeker & Escobar, 1998). Therefore, all correlation comparisons are conducted with latent variable modeling technique (i.e., SEM) because the procedure automatically controls for scale unreliability. We used raw item scores as observed indicators in the analysis. However, results based on uncorrected correlations gave the same conclusion.

We started by testing the nomological network of self-esteem. In an SEM model, a positive self-esteem latent factor was created from positive self-esteem items and a negative self-esteem latent factor was created from negative self-esteem items. To compare their correlation with an external variable such as full-scale conscientiousness, a third latent factor was created from conscientiousness items. All three latent factors, i.e., positive self-esteem, negative self-esteem, and full-scale conscientiousness, were constrained to have a variance of unity, and they were allowed to covary with each other. After this initial model, a follow-up model was set up to constrain the correlations of positive and negative self-esteem with full-scale conscientiousness to be equal. Scaled difference chi-square tests (Satorra & Bentler, 2001, 2010) were then conducted to compare the initial model and the follow-up model. A statistically significant difference indicates that positive and negative self-esteem correlate differently with full-scale conscientiousness. We first compared the correlations of positive and negative self-esteem with a full-scale variable (e.g., full-scale conscientiousness), followed by a positively valenced variable (e.g., positively valenced conscientiousness) and finally a negatively valenced variable (e.g., negatively valenced conscientiousness).

For Sample 1, neither positive nor negative self-esteem had much advantage over each other in terms of their correlations with other full-scaled constructs (see the top left panel of Table 2). Only one of the eight correlation comparisons favored positive self-esteem. However, a different picture emerged when correlations with positively and negatively valenced scale scores were compared. For positively valenced scale scores, four out of eight cases favored positive self-esteem and none favored negative self-esteem (see the middle left panel). In contrast, for negatively valenced scale scores, four comparisons favored negative self-esteem and none favored positive self-esteem (see the bottom left panel). Therefore, Sample 1 showed that positive self-esteem scores tend to correlate more strongly with positively valenced scores than do negative self-esteem scores. Conversely, negative self-esteem scores tend to correlate more strongly with negatively valenced scores than do positive self-esteem scores. The results for Sample 2 (see the right panel) were similar to the results for Sample 1.

Our results demonstrated that item valence can affect a researcher’s substantive conclusion regarding the dimensionality of self-esteem. We further examined the generalizability of this finding with the construct extraversion and we found a similar pattern of the results (see Supplementary Table S7).

Variance Explained by Item Valence

One may question whether the effect of valence is large enough to be a legitimate concern. The MTMM analyses allowed us to calculate the magnitude of the overall construct effect and the overall valence effect in $M_{2\ \text{valence}}$ in two samples (Schmitt & Stults, 1986). The two valence factors accounted for 9.69% of the variance in the observed scores in Sample 1 and 7.98% of the variance in the observed scores in Sample 2. By comparison, the construct effect accounts for close to 50% of the variance in the observed scores (45.98% in Sample 1; 49.78% in Sample 2). Thus, the amount of variance attributable to valence is approximately one fifth of the amount attributable to the construct and is by no means trivial.
To examine the nature of the valence effects estimated in the MTMM procedures described above, we included two latent factors of social desirability response style (i.e., impression management and self-deception) in the Sample 2 CTCM model, and these latent factors were allowed to correlate with all construct factors and the two valence factors. The two valence factors remain orthogonal to other construct factors (see Figure 1). For each of the two social desirability response style factors, we were able to create three parcels for positive items and three parcels for negative items. All latent factors have a variance of unity, and all factor loadings were freely estimated. Again, we used Steyer and MacCallum’s (2010) procedure to create 100 randomly parceled data sets and estimate the average of parameter values among these data sets. The results are shown in Table 3. Overall, social desirability response style correlated positively with the positive-valence factor and negatively with the negative-valence factor, meaning that the two extracted method factors are associated with social desirability response styles, although their correlations are very weak. Recent researchers have suggested that the relationship of social desirability response style with other constructs may be nonlinear (Borkenau, Zaltauskas, & Leising, 2009; Dunlop, Telford, & Morrison, 2012); however, examination of bivariate scatterplots in the current study did not show apparent nonlinear relationships between social desirability and the item valence factors. Overall, social desirability response style can only partially explain the valence effect.

### General Discussion and Recommendations

Our in-depth investigations of item valence contribute to the literature in at least three significant ways. First, our most important finding is that valence can affect not only factor analysis, but also the magnitude of the correlations among
constructs in nomological network analyses. If, as is often the case (e.g., Credé et al., 2009), the scales included in these analyses are measured with uniformly valenced items, differences in the pattern of correlations cannot be unambiguously interpreted as evidence for the distinction between focal constructs (e.g., between positive and negative self-esteem). Second, some researchers have found that reverse-keyed items do not load on the same factor as regular-keyed items and consequently recommended the exclusion of reverse-keyed items in construct measurement (e.g., Schriesheim, Eisenbach, & Hill, 1991). This advice should be taken with caution as our results clearly demonstrate that such a research practice will lead to bias in construct correlations. Finally, we demonstrated that the observed effects of valence cannot be fully explained by social desirability response bias and therefore should be addressed as a separate issue in measure development and evaluation. We discuss these three issues in detail in the following subsections.

Implications for Construct Dimensionality Research

Factor Analysis

The current findings have substantial implications for construct dimensionality research. Factor analysis is based on the assumption that respondents will give similar answers to items that measure a common underlying factor (e.g., extraversion), but this analysis does not account for other non-substantive factors that affect item correlations (e.g., restriction of range, item distribution properties, and item extremity; see Guilford & Fruchter, 1978; McPherson & Mohr, 2005). The current research has demonstrated that item valence is another non-substantive factor that can influence the results of simple factor analytic procedures (such as exploratory factor analytic models and basic CFA models without method factors). A more advanced set of techniques, MTMM CFAs, is surprisingly underutilized (Marsh, Scalas, & Nagengast, 2010). MTMM CFAs are extensions to the common CFA model and can be used in popular structural equation modeling (SEM) programs (e.g., Amos, Lisrel, EQS, Mplus). We strongly recommend that researchers use this technique to check for the valence effect (see Eid and Diener, 2006, for a review of these models).

Nomological Network Analysis

To ensure a fair comparison in correlation coefficients, researchers might consider using measures with a balanced set of opposite-valenced items. Our empirical results demonstrate that a measurement instrument with a balanced set of positively valenced and negatively valenced items is least likely to show differential correlations with the opposite pole of a construct (e.g., extraversion vis-a-vis introversion). In practice, we realize that it is difficult to always use measures with a balanced set of oppositely valenced items because many psychological scales consist of predominantly regular-keyed measurement items. One solution to this problem is to re-weight scale items so the positively and negatively valenced items have the same overall contribution to a construct’s final score. For instance, if extraversion is measured by six positively and three negatively valenced items, researchers can give twice as much weight to the negatively valenced items. However, this method cannot be implemented on measures without any positively valenced items or any negatively valenced items.

A long-term solution to the item valence problem is to formulate items that are low in evaluative content (Bäckström, Björklund, & Larsson, 2009), in addition to creating measurement instruments that contain a balanced set of oppositely valenced items. Recently, a notable study by Bäckström et al. (2009) discovered that items in Big Five personality inventories are saturated with evaluative content that causes personality factors to correlate with each other even though they are theoretically orthogonal. When these researchers minimized item valence by reframing the personality items to be more neutral in meaning, the correlations among personality factors were substantially weaker, although they did not disappear entirely. These results suggest that the common variance among personality factors comes partially from par-

---

4Some researchers may question whether our item valence effect is simply caused by item extremity (Spector, van Katwyk, Brannick, & Chen, 1997). Extreme items may load on two separate factors (McPherson & Mohr, 2005). However, this alternative reason does not explain two of our important findings, namely (a) why items of the same valence across different measures will load on the same method factor in our MTMM analysis, and (b) why constructs of the same valence (e.g., introversion and neuroticism) will correlate stronger with each other (as compared to constructs of opposing valence). In addition, McPherson and Mohr (2005) found that even for the regular-keyed and reverse-keyed items that are moderate in wordings, they still load on two separate factors in factor analysis. Item extremity thus cannot fully explain our findings.
participants’ sensitivity to item valence. Their research did not investigate the ramifications of item valence on construct dimensionality debates as was done in the present research. By minimizing item valence during the scale development process, researchers can attenuate the inflated variance that is common among the items and among the constructs.

Another way to prevent the item valence effect, according to Campbell and Fiske (1959), is to maximize method heterogeneity among measures used in a nomological network analysis. For example, to test for the nomological network of optimism and pessimism, in addition to using a self-report measure of stress or well-being as potential correlates, researchers could include objective measures in the analysis. Physiological measures such as heart rate or cortisol level are not subject to the valence effect. Scores based on observable behaviors such as students’ absenteeism are also not prone to this bias. If differences in the correlations are observed for the self-report measures but not the physiological or behavioral measures, it would suggest that the former might be due to valence effects. Only in cases where differences are consistently observed in measures uncontaminated by the valence effect can a strong case be made for the bi-dimensionality of the focal construct. When a sample size is large, researchers may employ MTMM techniques to control for method-specific effects when examining the network of associations among constructs of interest.

Use of Other Strategies

The evidence provided in this study shows that valence effects can contribute to the apparent bi-dimensionality of a construct. However, unambiguously testing the dimensionality of a construct with factor analysis and nomological network analyses is difficult, because both analyses are correlation- or covariance-based techniques that are liable to the item valence effect. Thus, researchers should consider other strategies. These strategies include experimental techniques (using a controlled setting to manipulate two focal constructs [e.g., optimism and pessimism] in distinct experimental conditions) and response process analyses (building psychometric models for rigorous testing of the cognitive processes underlying participants’ item response behaviors). Detailed elaboration on these techniques is provided elsewhere (Borsboom & Mellenbergh, 2007; Borsboom, Mellenbergh, & van Heerden, 2004).

Exclusion of Reverse-Keyed Items

Previous authors have suggested excluding reverse-keyed items in psychological instruments because these items contribute to measurement artifacts (e.g., Lindwall et al., 2012; Magazine, Williams, & Williams, 1996; Schriesheim et al., 1991; Schriesheim & Eisenbach, 1995; van Sonderen et al., 2013). Our finding, in contrast, illustrates that this practice can result in distortion of correlations, leading to mis-leading conclusions regarding the substantive issues under investigation. Although it is increasingly popular to exclude reverse-keyed items in construct measurement, we do not advocate this practice. As noted by a reviewer, excluding reverse-keyed items is also a step backward from past research because a scale without reverse-keyed items cannot mitigate acquiescence response bias (Jackson & Messick, 1958), or participants’ tendency to endorse an item regardless of its content.

Social Desirability and Potential Explanations for the Valence Effect

Our results suggest that the social desirable response bias only partially explained the valence effect. Therefore, there must be other mechanisms involved. One possibility is suggested by Showers’s (1992) research showing that some individuals have a tendency to organize positive and negative knowledge into separately valenced memory categories. Credé et al. (2009) similarly argued that, when confronted with positively valenced items, respondents are likely to tap into positive memories that justify agreement. Similarly, when confronted with negatively valenced items they tap into negative memories that can lead to agreement. This tendency to focus on valence-relevant memories leads to an increase in the correlations among similarly valenced items and reduces the correlation among opposite-valenced items. As a second potential explanation, some researchers have suggested that individuals may have ambivalent or dialectical thinking patterns (Peng & Nisbett, 1999; Spencer-Rodgers, Boucher, Mori, Wang, & Peng, 2009), characterized by the belief that logically contradictory statements (such as “I am extroverted” and “I am introverted”) can both be true. These individuals have been described to have “low internal consistency of the content of one’s global self-beliefs or ‘personality’” (Spencer-Rodgers et al., 2009, p. 30). As a result of their lack of a fixed view of themselves and the world, they agree or disagree simultaneously with what to others appear to be contradictory statements. We encourage further research to examine these and other possible causes of item valence.

Limitations and Future Directions

Like most psychological research, the current study has limitations. First, because we used only university students as respondents, we cannot necessarily generalize our results to other populations. However, although affirmation of our results using samples with different socioeconomic status might be beneficial, there is no evidence suggesting that participants’ response styles differ substantially between a college sample and other samples. Second, although we have included a wide range of personality and social psychology measures in our empirical investigation, it would be beneficial to include more measures (e.g., work engagement) to further examine the generalizability of our findings. Third,
we employed item parcels in our structural equation modeling analysis due to the large number of items in the current study (up to the maximum of 114 items in a model). Although we have adopted the latest suggestion by Sterba and MacCullum (2010) to control for parcel-allocation variability, item parceling may not be the ideal method because it obscures the influence of individual item characteristics (e.g., item residual covariances). Moreover, participants may differ in their perceived favorability of an item (Messick & Jackson, 1961) and this has not been considered in the current study. Future research might incorporate individual differences in item valence when modeling MTMM data. Finally, the current research has not fully explored the nature of the valence effect. The main purpose of this paper was to discover how item valence affects dimensionality debates in general, rather than to theorize or test all potential mechanisms underlying the valence effect. Nevertheless, understanding the nature of item valence is extremely important in advancing our current knowledge of how individuals interpret and respond to survey items (Borsboom et al., 2004), and we encourage future researchers to explore this question.

Conclusion

The major goal of our paper was to demonstrate that item valence has a strong potential to influence conclusions with regard to the dimensionality of a construct. Our findings go beyond the well-established item-keying effect in factor analysis to demonstrate that regular and reverse-keyed items can differ in valence, and that this valence difference is at least partially responsible for the emergence of multiple factors. More importantly, we demonstrate that the item valence can also influence the correlation between measures in nomological network analyses, and in correlational research in general. We demonstrate that this valence effect is only partially explained by social desirability response bias, and suggest other mechanisms to be investigated in future research. An important practical implication of our findings is that they draw attention to problems that can result from the increasingly common practice of excluding reverse-keyed items.

SUPPLEMENTAL DATA

Supplemental data for this article can be accessed on the publisher’s website.

ARTICLE INFORMATION

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant Numbers SRG2013-00050-FED and MYRG2015-00076-FED from the University of Macau to the first author, and Grant Number 410-1020-1923 from the Social Sciences and Humanities Research Council to the second author.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like thank Sampo Paunonen, Richard Neufeld, Richard Goffin, Heather Laschinger, Mitch Rothstein, Irene Cheung, Yanjun Guan, Yuan Wan Ho, Etienne LeBel and Thomas O’Neill for their contribution.

REFERENCES

Altemeyer, R. A. (1981). Right-wing authoritarianism. Winnipeg, Canada: University of Manitoba.

Bäckström, M., Björklund, F., & Larsson, M. R. (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. doi: 10.1016/j.jrp.2008.12.013

Beauvois, J. L., & Dubois, N. (2000). Affordances in social judgment: Experimental proof of why it is a mistake to ignore how others behave towards a target and look solely at how the target behaves. Swiss Journal of Psychology, 59, 16–33. doi: 10.1024/0142-0185.59.1.16

Bentler, P. M., & C.-P. Chou (1987). Practical issues in structural modeling. Sociological Methods & Research, 16, 78–117. doi: 10.1177/00491241418701601004

Borkenau, P., Zaltauskas, K., & Leising, D. (2009). More may be better but there may be too much: Optimal trait level and self-enhancement bias. Journal of Personality, 77, 825–858. doi: 10.1111/j.1467-6494.2009.00566.x

Borsboom, D., & Mellenbergh, G. J. (2004). The concept of validity. Psychological Review, 111, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81–105. doi: 10.1037/h0046016
Rosenberg, S., & Olshan, K. (1970). Evaluative and descriptive aspects in personality perception. *Journal of Personality and Social Psychology, 16*, 619–626. doi: 10.1037/h0030081

Rubin, Z., & Peplau, L. A. (1975). Who believes in a just world? *Journal of Social Issues, 31*, 65–90. doi: 10.1111/j.1540-4560.1975.tb00997.x

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514. doi: 10.1007/BF02296192

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*, 243–248. doi: 10.1007/s11336-009-9135-y

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*, 1–22. doi: 10.1177/014662168601000101

Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement, 51*, 67–78. doi: 10.1177/00131644911511005

Schriesheim, C. A., & Eisenbach, R. (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. doi: 10.1177/014920639502100609

Showers, C. (1992). Compartmentalization of positive and negative self-knowledge: Keeping bad apples out of the bunch. *Journal of Personality and Social Psychology, 62*, 1036–1049. doi: 10.1037/0022-3514.62.6.1036

Sidanius, J., & Pratto, F. (2001). *Social dominance: An intergroup theory of social hierarchy and oppression*. New York, NY: Cambridge University.

Spector, P. E., van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don’t reflect two constructs: How item characteristics can produce artifactual factors. *Journal of Management, 23*, 659–677. doi: 10.1177/014920639702300503

Spencer-Rodgers, J., Boucher, H. C., Mori, S. C., Wang, L., & Peng, K. (2009). The dialectical self-concept: Contradiction, change and holism in East Asian cultures. *Personality and Social Psychology Bulletin, 35*, 29–44. doi: 10.1177/0146167208325772

Sterba, S. K., & MacCallum, R. C. (2010). Variability in parameter estimates and model fit across random allocations of items to parcels. *Multivariate Behavioral Research, 45*, 322–358. doi: 10.1080/00273171003680302

Tomás, J. M., & Oliver, A. (1999) Rosenberg’s self-esteem scale: Two factors or method effects. *Structural Equation Modeling, 6*, 84–98. doi: 10.1080/10705519909540120

Vagg, P. R., Spielberger, C. D., & O’Hearn, T. P., Jr. (1980). Is the state-trait anxiety inventory multidimensional? *Personality and Individual Differences, 1*, 207–214. doi: 10.1016/0191-8869(80)90052-5

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let’s learn from cows in the rain. *PLoS ONE, 8*. doi: 10.1371/journal.pone.0068967

Widaman, K. F. (1985). Hierarchically nested covariance structures models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1–26. doi: 10.1177/014662168500900101

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Philippines: Addison-Wesley.

Yuan, K.-H., Bentler, P. M., & Kano, Y. (1997). On averaging variables in a CFA model. *Behaviormetrika, 24*, 71–83. doi: 10.2333/bhmk.24.71