Running Head: Response Heterogeneity Between Positively and Negatively Worded Items

Investigating Response Heterogeneity in the Context of Positively and Negatively Worded Items by Using Factor Mixture Modeling

Chester Chun Seng KAM *

University of Macau, China

Xitao FAN

The Chinese University of Hong Kong (Shenzhen), China

* Corresponding Author:

Chester Chun Seng KAM

Faculty of Education

University of Macau

Macau, China

Email: ChesterKam@umac.mo

Kam, C. C. S., & Fan, X. (accepted). Investigating response heterogeneity in the context of positively and negatively worded items by using factor mixture modeling. *Organizational Research Methods*. https://doi.org/10.1177/1094428118790371. Reprinted by permission of SAGE Publications.

ABSTRACT

Factor mixture modeling was used to investigate potential response incongruity between positively and negatively worded items. Survey respondents (N = 591) answered questions about job satisfaction and dissatisfaction. Results revealed two classes of respondent: a majority class, who generally do not have problems answering positively and negatively worded items; and a minority class, who have serious trouble with negatively worded items. With the exclusion of the minority class, job satisfaction and dissatisfaction were found to be essentially unidimensional, rather than bidimensional as previous research had suggested. These results not only challenge previous findings regarding the bidimensionality of job satisfaction, but also question the widespread research practice of assuming population homogeneity in survey responses. A flow diagram illustrating the analytic procedure and an Mplus syntax program are provided so that researchers can conduct similar investigations on constructs of interest.

Keywords: item wording; factor mixture modeling; unidimensionality

Investigating Response Heterogeneity in the Context of Positively and Negatively Worded Items by Using Factor Mixture Modeling

Whereas positively worded items measure the presence of a construct, negatively worded items measure its absence. Although positively and negatively worded items are supposed to measure the same construct, the correlation between them is far from perfect: indeed, it is well known that survey respondents may answer positively worded and negatively worded items differently. This result has led to the "dimensionality debate" concerning many constructs, such as presence versus absence of anxiety (Vautier & Pohl, 2009), optimism versus pessimism (Kam & Meyer, 2012), positive versus negative self-esteem (Marsh, Scalas, & Nagengast, 2010), and job satisfaction versus dissatisfaction (Credé, Chernyshenko, Bagraim, & Sully, 2009).

Consider job satisfaction. Some researchers argue that satisfaction (positively worded) items and dissatisfaction (negatively worded) items are the opposite ends of a unidimensional construct (e.g.,Kam & Meyer, 2012; Rauch, Schweizer, & Moosbrugger, 2007), whereas others believe that they represent separate, distinct constructs (e.g., Herzberg, Glaesmer, Hoyer, 2006; Marshall et al., 1992). A common yet untested assumption in most of these studies is that respondents approach survey items in a similar manner and thus the same dimensionality solution applies to everyone in the data set.

The purpose of the current article is twofold: first, to demonstrate the untenability of this 'one-size-fits-all' assumption, and second, to propose a method researchers can use to test this assumption. The paper is organized as follows. We begin by discussing why researchers include negatively worded items in construct measurement, and why respondents may answer positively and negatively worded items differently. We then propose a method to investigate the assumption, and finally demonstrate the utility of the method by describing an empirical investigation.

Negatively Worded Items

Psychometricians have long been concerned about acquiescence response style, that is, participants' tendency to agree with survey items regardless of content (Nunnally & Bernstein, 1994). When survey scales include items worded in the same keying direction (e.g., for job satisfaction, asking

'Are you satisfied with your job?'), construct mean scores can be systematically inflated due to the effect of response style. In addition, acquiescence response style can affect construct correlations (Kam & Meyer, 2015a). If acquiescence affects any two constructs, correlations between them will be biased in the positive direction: positive correlations will be inflated (i.e., increase in the absolute magnitude of a positive correlation) and negative correlations deflated (i.e., decrease in the absolute magnitude of a negative correlation).

To control for acquiescence response style, researchers typically include reverse-keyed items[1] (e.g., 'Are you dissatisfied with your job?'). Reverse-keyed items have other advantages as well (Weijters & Baumgartner, 2012): (a) they widen the sampling of a construct's domain, thus leading to higher content validity in the measurement of the construct (Tourangeau, Rips, & Rasinski, 2000); and (b) they act as a kind of cognitive speed bump for respondents who are trying to breeze through a survey (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), thus potentially encouraging them to pay closer attention. Therefore, psychometricians often recommend using a balanced set of positively and negatively worded items to minimize the effect of acquiescence response style, and to enhance the validity of a construct's score (Nunnally & Bernstein, 1994).

Assuming that acquiescence affects positively and negatively worded items equally, the effect of the response style will cancel out, and the resulting score is potentially more valid than when a construct is measured only by positively worded items (see Kam & Meyer, 2015a, 2015b, for more advantages of using both positively and negatively worded items).

Despite the apparent advantages of summing up scores from positively and negatively worded items, some researchers have warned against the use of negative items. They pointed out that even though positively and negatively worded items were supposed to measure the same construct, the two types

[1] The current article uses the terms regularly-keyed items and positively worded items interchangeably. Both measure the presence of a construct. Similarly, the current article uses the terms reverse-keyed items and negatively worded items interchangeably. Both measure the absence of a construct.

might not be highly correlated. In the realm of job satisfaction, for example, satisfaction (positively worded) items did not correlate highly with dissatisfaction (negatively worded) items (Credé et al., 2009). Similarly, for optimism, optimism items did not correlate highly with pessimism items. Some researchers took this as evidence of construct bidimensionality: positively and negatively worded items were measuring different constructs (Dalbert, Lipkus, Sallay, & Goch, 2001; Credé et al., 2009; Fincham & Linfield, 1997). These researchers suggested to calculate two separate scores from positively and negatively worded items and to treat them as distinct constructs. Other researchers believed less-than-perfect correlations were an artifact caused by participants having trouble responding to negatively worded items (Carmines & Zeller, 1979; Greenberger et al., 2003; van Sonderen, Sanderman, & Coyne, 2013; Vautier & Pohl, 2009). They suggested not to use negatively worded items to measure a construct.

Response Incongruity between Positively and Negatively Worded Items

Negatively worded items have been found to have lower item quality than positively worded ones (Holden & Fekken, 1990; Schriesheim, Eisenbach, & Hill, 1991; Schriesheim & Hill, 1981; Sliter & Zickar, 2014), suggesting that participants may have more problems responding to negatively worded items. For example, Sliter and Zickar (2014) showed that negatively worded items had lower item discrimination than positively worded ones, implying that negatively worded items correlated less well with their latent construct. Holden and Fekken (1990) found that participants, over a period of one month, were less consistent in their responses to negative items than to positive ones. This result could mean that negative items are more vulnerable to transient or situational influences than positive items. Holden and Fekken also found that on a clinical scale, negatively worded items correlated slightly but *negatively* with clinician ratings, implying that negative items have lower criterion validity than positive items. Schriesheim and colleagues revealed that negatively worded items had substantially lower internal consistency reliability than positively worded items, implying that participants provided less consistent responses to negatively worded items (Schriesheim & Hill, 1981; Schriesheim et al., 1991).

A potential limitation of these studies, however, is that they assumed population homogeneity: all respondents are considered to be a sample from the same population. It is possible, though, that only

some—not all—of the respondents have trouble with negatively worded items. Therefore, any research findings based on the assumption of population homogeneity is potentially misleading, if not fundamentally flawed.

There is some evidence that individual differences may influence the observed relationship between positively and negatively worded items. Some researchers have found that response inconsistency is correlated with certain personality constructs (de Jonge & Slaets, 2005; Michaelides, Koutsogiorgi, & Panayiotou, 2015; Quilty, Oakman, & Risko, 2006). Inconsistency was also found to be associated with low cognitive ability (Gnambs & Schroeders, 2017), with low reading ability (Corwyn, 2000; Marsh, 1996), and with trait anxiety (Tomás et al., 2013). Therefore, it is possible that the relationship between positively and negatively worded items, and thus the observed dimensionality of a construct, differs across individuals in a population. However, there is a dearth of research on potential population heterogeneity as such.

Recovering Population Heterogeneity

Factor mixture modeling (FMM) has been employed to investigate population heterogeneity (e.g., Lubke & Muthén, 2005; Raykov, Marcoulides, & Chang, 2016), and this approach may be employed to investigate population heterogeneity as related to positively and negatively worded items. In many statistical analyses, it is assumed that respondents all come from the same underlying population, hence population *homogeneity*. The possibility of population *heterogeneity* is not examined in many situations, partially due to the limitations of many analytical approaches. Usually, testing for population heterogeneity is possible only when researchers have prior knowledge regarding participants' characteristics. This is the case with between-group comparisons in *t*-tests or multi-group structural equation modeling (SEM) analysis. In a multi-group SEM, for example, Asian and White participants can be compared if information regarding ethnic group membership is available. In that case, the heterogeneity of Asian and White participants becomes an empirical question. Put differently, these analyses are only applicable when the information concerning the membership (e.g., Asian vs. White, male vs. female) is already known (i.e., observed).

In many situations, population heterogeneity may exist, but this is unknown. Similarly, the membership information concerning the possible sub-populations is also unknown (i.e., latent). An advantage of FMM is that it does not require researchers to have prior knowledge regarding group membership; instead, based on the data of the response variables, the algorithm searches for qualitatively distinct subgroups within the sample data. If identified, members of the same subgroup are qualitatively more similar to each other, but more different from those in other subgroups. FMM is thus a useful tool to explore patterns of hitherto unobserved heterogeneity.

Purpose of the Current Study

The purpose of the current study is to investigate possible incongruity in responses to positively and negatively worded items. As we have noted, previous studies have assumed population homogeneity in participants' response patterns. However, there is growing evidence of possible population heterogeneity in these situations. For example, in the area of research on organizational commitment, employees could be in qualitatively distinct subgroups, each with its own validity, antecedents, and outcomes (Kam, Morin, Meyer, & Topolnytsky, 2016). The same phenomenon (i.e., distinctly different subpopulations) may with respect to participants' responses to positively and negatively worded items. As noted earlier, individual differences in answering positively and negatively worded items have been identified (de Jonge & Slaets, 2005; Gnambs & Schroeders, 2017; Michaelides et al., 2015; Quilty et al., 2006). Therefore, the assumption of homogeneity in this specific situation needs to be investigated.

The current study is thus a timely investigation about the question of potential population heterogeneity in responses to positively and negatively worded items. As a preliminary study, we focus on a construct the dimensionality of which is currently a matter of some debate: job satisfaction. Although some past research did not show strong evidence for differential validity between satisfaction (positively worded) items and dissatisfaction (negatively worded) items (e.g., Hines, 1973), more recently, some research challenged this finding. In particular, Credé et al. (2009) found that job satisfaction and dissatisfaction items were imperfectly correlated, implying that the construct could be bidimensional.

Similar to other researchers, Credé et al. made an assumption of population homogeneity (i.e., the bidimensionality of job satisfaction applies to everyone in the sample).

The current study investigates this assumption. If, for instance, a subgroup of respondents would have difficulty answering negatively worded items, then the overall correlation between satisfaction and dissatisfaction would be less than perfect, which could be mistakenly construed as evidence of bidimensionality for the entire group. If this reasoning is correct, exclusion of the subgroup that had difficulty in responding to negatively worded items should better recover the 'actual' correlation between satisfaction and dissatisfaction.

Although the current study focuses on job satisfaction, we would like to make it clear that we are using the construct of job satisfaction only as an example. Our aim is to propose a general analytical procedure for applied researchers, so that they can investigate the issue dimensionality of any construct of interest when positively and negatively worded items are involved.

Methods

Participants

The original sample was 828 full-time working respondents from the U.S. and Canada recruited by an online data collection company. For the purpose of screening out careless respondents, the survey included five instructed response questions that asked participants to select a particular response (e.g., 'strongly disagree') or skip an item. Previous research (Kam & Meyer, 2015a; Meade & Craig, 2012) has shown that instructed response questions are effective in screening out careless respondents. Kam and Meyer (2015a) also showed that careless responding can severely distort the correlation between positively and negatively worded items; as a result, screening out careless respondents could improve data quality. These findings were supported by other research (e.g., Huang, Liu, & Bowling, 2015). The final sample after screening out careless respondents[2] included 591 respondents (257 male and 334 female; M_{age} = 42.53, SD_{age} = 10.56), who answered all the instructed questions correctly. Although other parts of

[2] For procedural details of screening out careless respondents, please see Kam and Meyer (2015a).

the data have been analyzed for other issues and published elsewhere (Authors), the findings reported here have not been previously published.

Instrument

We designed five pairs of overall job satisfaction items by adapting existing items from the literature. Two items in a pair are parallel in wording. For example, a satisfaction item was 'Considering everything, I am satisfied with my job,' and the corresponding dissatisfaction item was 'Considering everything, I am dissatisfied with my job'. Parallel items ensure that the only difference between them was item wording direction (i.e., satisfied vs. dissatisfied). All items were worded such that participants were encouraged to consider their overall impression about their job, as opposed to a specific aspect. Previous research (Wanous, Reichers, & Hudy, 1997) has found that a custom-made, one-item job satisfaction item has considerable validity, and therefore our five-item scale should be an even better measure of the construct. The items are shown in Supplementary Material S2.

To facilitate the interpretation of the results, all satisfaction and dissatisfaction items were coded such that a higher value represents higher satisfaction. Therefore, when job satisfaction correlated perfectly with job dissatisfaction, their latent correlation would be +1 (rather than -1) in the current study.

Analytical Procedure

The data analyses were done by using the factor mixture modeling program under Mplus 7.11 (Muthén & Muthén, 1998-2017). The procedural steps for testing response incongruity between positively and negatively worded items are shown in Figure 1. These procedural steps are similar to those used to examine the Trait × Method interaction under the MTMM framework proposed by Litson et al. (2017). We will discuss the major differences between them in Discussion section later.

The fit of the single-class configural invariance models is first examined (Step 1). Positively worded (job satisfaction) and negatively worded (job dissatisfaction) items loaded on two separate factors, and the factors were allowed to covary with each other (Figure 2). One reference item was chosen for positively worded items and another for negatively worded items to have the loading of 1. These

reference items had item intercepts of 0 so that factor means could be estimated. This can be observed with the following model specifications (Litson et al., 2017):

$$y_{1jk} = \tau_{1jk} + \varepsilon_{1jk}\text{, when the reference item is item 1 for method } j \text{ } (i = 1)$$

$$y_{ijk} = \alpha_{ijk} + \tau_{ijk} + \varepsilon_{ijk}\text{, for other items}$$

where y refers to the observed item score, τ refers to true construct scores, i refers to item number, j refers to a specific method, and k refers to class number from 1 to c (when a multi-class model is specified).

Factor means (τ) can be estimated from the mean score of the reference item (y_{1jk}). In addition, when a scale is comprised of items with parallel wording, failure to account for covariance due to item parallel wording will bias trait variance (Cole, Ciesla, & Steiger, 2007). Therefore, residual covariances were added between each pair of parallel items. If the fit of this model was good, we tested a more parsimonious model of weak invariance in which loadings of parallel items between positively and negatively worded items were constrained to be identical. Weak invariance is necessary to test equality of factor variances between satisfaction and dissatisfaction. If the fit did not deteriorate significantly, we tested an even more parsimonious model of strong invariance, in which item intercepts were constrained to be identical between parallel items. Strong invariance (i.e., equal factor loadings and item intercepts) is necessary to test equality of factor means between satisfaction and dissatisfaction. If the fit did not significantly worsen, we examined the strict invariance model in which factor loadings, item intercepts, and item residual variances were all constrained to be equal between parallel items. Comparisons among these nested models were made using the Satorra-Bentler chi-square difference statistic (Satorra, 2000) because robust maximum likelihood (MLR) estimator was used for data that deviates from multivariate normality. A non-significant chi-square statistic means that a more constrained model (e.g., weak invariance) does not fit significantly worse than a less constrained model (e.g., configural invariance), and thus the more constrained model is preferred.

After establishing the best single-class model, we proceeded to test a model with one additional class (Step 2). Using factor mixture modeling (FMM), we specified the number of classes to be two. The general specification of the model was the same as the single-class configural invariance model, except

that model parameters were allowed to differ across classes. The factor loading of the reference items was still constrained to be 1 in both classes. If the baseline model fitted well, we proceeded to examine model parameters to test the possibility of invariance models (i.e., weak, then strong, then strict invariance) between methods within the same class (Step 2) and between classes (Step 3). We tested measurement invariance between methods within each class first, and then tested measurement invariance between classes, because our main purpose is to test comparability between positively and negatively worded items. Testing of measurement invariance between classes is a secondary concern, simply to derive a more parsimonious model.

Model comparisons were facilitated by referencing Akaike's information criterion (AIC) and Bayesian information criterion (BIC). A model with lower AIC and BIC values is preferred. Previous research has shown that BIC outperforms AIC in class identification (cite). Although not used for class selection, entropy indicates certainty of class assignment among respondents, with a higher value potentially indicating better accuracy in this process. After identifying the best model in the two-class models, we tested three-class models with one additional class. The process of adding additional classes stopped when all invariance models did not converge or when the best higher-class model fitted worse than the best model with one less class. Equality of variance, factor means, and factor correlations were considered after the best model was identified (Step 4). Previous research suggested the possibility of response inconsistency among positively worded items or among negatively worded items. When inconsistency was found in at least one class, we examined respondents' raw responses to positively or negatively worded items (Step 5).

Results

Single-class Models

Configural invariance model (CMI_1) fit the data well ($\chi^2 = 82.81$, $df = 29$, $p < .001$, TLI = .97, CFI = .96, RMSEA = .06, SRMR = .04). All factor loadings were sizeable, with standardized βs ranging from .69 to .85 (detailed results not shown) and all statistically significant. The factor correlation between job satisfaction and dissatisfaction was estimated to be .76 (95% C.I. = [.70, .82]), suggesting

bidimensionality. Two of the five-item residual covariances were statistically significant (.13 and .18 respetively, *p*s < .03), suggesting a weak effect of parallel item wording. A more parsimonious model of weak invariance fit the data statistically worse ($\Delta\chi^2 = 18.24$, $\Delta df = 4$, $p = .001$), indicating that item factor loadings differ between job satisfaction items and dissatisfaction items. Due to the much-less-than-perfect factor correlation (i.e., .76) and the metric non-invariance, applied researchers may conclude that job satisfaction and dissatisfaction are distinct yet related constructs, with positively and negatively worded items demonstrating non-comparable qualities that should not be summed together to form a total score.

Two-class Models

The two-class configural invariance model (CMI_2) showed an improvement in fit compared to the one-class model, as indicated by the former's lower BIC and AIC values (see Table 1). The results pointed to possible population heterogeneity, and thus it may not be appropriate to assume that all respondents came from the same population, as previous research on job satisfaction typically did. The data showed a majority class (with ~78% of respondents) and a minority class (~22%).

The majority class had similar values of unstandardized factor loadings across the items of job satisfaction and those of dissatisfaction, with reasonably small standard errors in these parameter estimates (*SE*s < .07). The second class, in contrast, featured fluctuating unstandardized factor loadings even within the item group of job satisfaction or that of dissatisfaction, with unreasonably large standard errors (*SE*s = 0.36 to 2.02). The large standard errors suggest that FMM has severe difficulties in parameter estimation in the minority class. Given the estimation uncertainty in this class, we reasoned that it would not be meaningful to test weak invariance between classes, as the establishment of weak invariance between classes could simply be the result of large standard errors in the second (minority) class. We will return to this issue of large standard errors in the minority class later. For now, we proceed to test measurement invariance between methods (i.e., positively vs. negatively worded items) in the first (majority) class.

The model with weak invariance between methods in the majority (WMI_2) class had lower BIC and AIC values, meaning that the new model fit the data better. The size of the class allotment did not

change much from the previous model. The best model, however, was the one with strong invariance between methods in the majority class, as it had even lower BIC and AIC values (Table 1). Further analysis showed that the strict invariance model ($StMI_2$) featured non-replicable likelihood value and improper solution (an abnormally large factor loading and an even larger standard error in one item in class 2). We also inspected the confidence interval for those unstandardized residual variances in the SMI_2 model. The residual variances were consistently larger for job dissatisfaction than for job satisfaction in the majority class, meaning that the between-method strict invariance model is likely untenable. Therefore, we concluded that the best two-class model was the model with strong between-method invariance in the majority class (SMI_2).

Three-class Models

The three-class configural invariance (CMI_3) model failed to produce replicable maximum likelihood values. The best maximum likelihood values also fluctuate wildly, suggesting substantial instability of the solutions. Therefore, we concluded that the best solution was the two-class solution, with strong between-method invariance in the majority class (i.e., Model SMI_2 in Table 1).

Testing for Invariance and Strict Unidimensionality between Methods

We tested equality of factor means and factor variance in the majority class in our SMI_2 solution. The higher BIC and AIC values in the mean equality model suggested that job satisfaction and dissatisfaction, although having seemingly comparable factor means (3.91 vs. 4.03), are still statistically different. The results mirror previous findings that respondents have a slightly stronger tendency to disagree with negatively worded personality items than to agree with positively worded personality items (Kam, 2017). As a result, personality scores measured by negatively worded items generally have higher means than those measured by positively worded ones.

The higher BIC and AIC values in the variance equality model also suggest that job dissatisfaction has significantly higher variance than job satisfaction (0.79 vs. 0.56). Finally, the confidence interval for the latent correlation between satisfaction and dissatisfaction does not include 1 (estimated 95% C.I. = .87 [.82, .92]). Results thus suggest that job satisfaction-dissatisfaction is not a

strictly unidimensional construct. Later, we will test whether job satisfaction-dissatisfaction was an essentially unidimensional construct in the majority class. Due to the exceptionally large standard errors in the minority class, again it would not be meaningful to test measurement invariance and equality of mean and variance in this class.

The final parameter estimates in the two-class strict invariance solution are shown in Table 2. The factor loadings of the majority class were all sizeable, and the correlation between job satisfaction and dissatisfaction was strong. In the minority class, in contrast, both the factor loadings and the construct correlation suffer from large standard errors. The large standard errors imply that there is possible problem in participants' response patterns, and thus we proceed with this analysis.

Testing Essential Unidimensionality

According to the recent discussion by some methodologists (e.g., Reise et al., 2016; Rodriguez, Reise, & Haviland, 2016), it is often unrealistic to assume strict unidimensionality (e.g., perfect correlation between job satisfaction and dissatisfaction items). Items are supposed to measure a wide domain of a construct, and restricting item content to a specific aspect of the construct potentially hinders validity (Kam & Meyer, 2015b). For example, including the positive items of a construct (e.g., emotional stability items) may lead to insufficient sampling on the negative end of the same construct (neuroticism items). In addition, measuring a construct with only one particular method (e.g., regular-keyed items only) conflicts with the idea of multimethod measurement (MTMM; Campbell & Fiske, 1959).

Therefore, instead of pursuing strict unidimensionality, researchers recently suggested that the goal should be 'essential' (Gu, Wen, & Fan, 2017; Reise et al., 2016) or 'approximate' unidimensionality (Raykov & Marcoulides, 2016). The idea is that a construct can be considered unidimensional if the common trait factor among all items accounts for most of the construct variance. Gu et al. (2017) suggested calculating an ECV index to examine the strength of the common factor in a bifactor model. Unfortunately, FMM does not allow the specification of a bifactor model to measure ECV while keeping the class membership of the respondents to be the same in our previous analysis. Therefore, to calculate

ECV, we extracted participants' class membership and specified a different multitrait-multimethod (MTMM) model for the majority class (i.e., class 1).

We set up a bifactor model in which all items load on an overall factor and negatively worded items load on a method-specific factor. The calculation of explained common variance (ECV) is as follows:

$$\text{ECV} = \frac{\sum_{i=1}^{10} \lambda_{gi}^2}{\sum_{i=1}^{10} \lambda_{gi}^2 + \sum_{i=6}^{10} \lambda_{si}^2}$$

where λ_g^2 = factor loading of the common latent factor, λ_s^2 = factor loading of the method-specific factor, i = specific item number, with positively worded items ranging from 1 to 5 and negatively worded items ranging from 6 to 10. Gu et al. (2017) suggested that if the common factor (e.g., overall job satisfaction) explains more than 75% of the trait variance (i.e., ECV ≥ .75), researchers could assume essential unidimensionality even though the construct is not strictly unidimensional.

We therefore extracted class memberships from FMM and fit a bifactor model. First, we continued to find metric equivalence between satisfaction and dissatisfaction items under the common trait factor in this bifactor model. Second, results showed that ECV is .87 in the majority class, meaning that the common trait factor explains about 87% of the variance while the method-specific factor explains about 13% of the variance.[3] Given the large trait variance as compared to the method variance in the majority class, the job satisfaction construct is not strictly unidimensional but can be regarded as essentially unidimensional. We did not examine ECV in the minority class because the problematic item response patterns in this class render the examination not meaningful.

[3] Note that these ECV estimates do not include variances from residual correlation, because the purpose is to compare trait variance and the negative-wording method variance. The ECV cutoff value recommended by Gu et al.'s (2017) simulation was also based on the comparison between trait variance and method variance without correlated residuals.

Examining Response Patterns

We examined the response pattern with two indices. The first index is the difference (*d*) between the maximum response value and the minimum response value within items of the *same wording* direction for each respondent *j*'s responses. We calculated one value (d_{js}) for job satisfaction items and another (d_{jd}) for job dissatisfaction items. A higher *d* value implies more inconsistency even among items with the same wording direction across an individual's responses. The second index is the inconsistency index (denoted as i_{js} for job satisfaction and i_{jd} for job dissatisfaction) for each respondent. Respondents have the value of 1 for i_{js} when they answer 'agree' or 'strongly agree' to a job satisfaction item but 'disagree' or 'strongly disagree' to another job satisfaction item. Otherwise, the respondent will have the value of 0. We conducted the same calculation for the i_{jd} value with job dissatisfaction items. The *i* indices help us evaluate the extent to which respondents give consistent answers to items with the same wording direction.

We compared the *d* indices between the majority and minority classes. Results showed a nonsignificant difference for d_{js}, 0.77 (*SD* = 0.81) vs. 0.71 (*SD* = 0.84), $t(197.04) = 0.74$, $p = .46$, Cohen's $d = 0.075$, effect size $r = .038$. For d_{jd}, however, the majority group had significantly smaller value than the minority group, 0.89 (*SD* = 0.92) vs. 2.73 (*SD* = 0.74), $t(247.31) = -23.74$, $p < .001$, Cohen's $d = -2.197$, effect size $r = -.739$. In other words, the difference between the two classes in response consistency mostly involves job dissatisfaction items. When we further analyzed the *i* indices for both classes (Table 3), again we discovered that the minority class has the most trouble in providing consistent responses to dissatisfaction items. In contrast, the percentage of respondents giving consistent responses to both satisfaction and dissatisfaction items appears to be much larger for the majority class. These results suggest that an obvious difference between the two classes was participants' ability to provide consistent responses among dissatisfaction items: respondents in the minority class do not provide consistent answers to dissatisfaction items.

Discussion

The current study applied FMM to examine the possibility of population heterogeneity in responses to positively worded and negatively worded items. We included only careful respondents because careless ones have been shown to bias construct correlation in the direction of favoring bidimensionality (Kam & Meyer, 2015a). Whereas previous researchers have tended to examine construct dimensionality with simple factor analytic models, we have empirically demonstrated that FMM was able to meaningfully identify two distinct groups of respondents. Each group shows distinct response behaviors to positively and negatively worded items. Most interestingly, measurement invariance between methods was found for the group that apparently has no major problem answering positively and negatively worded items. A flow diagram delineates the process of data analysis (Figure 1), and two syntaxes with detailed explanations are provided for applied researchers who are interested in conducting similar analyses (please refer to Supplementary Material S1 and S2). In the present article, we have illustrated the utility of this method to help resolve the dimensionality debate of a popular construct (job satisfaction).

The debate regarding the dimensionality of job satisfaction started decades ago (Herzberg et al., 1957) and has not been entirely resolved. Employing exploratory factor analysis, researchers have found that a two-dimensional solution fits the data better than a unidimensional one (Credé et al., 2009). At the same time—and given that job satisfaction has an affective component—theorists have recently suggested that satisfaction possibly differs from dissatisfaction, mirroring the distinction between positive and negative affect in the emotion literature (Judge, Weiss, Kammeyer-Mueller, & Hulin, 2017). Our results, however, do not show strong support for the two-dimensional interpretation. Instead, we discovered a class of respondents, a little over 20% of them, who mostly give inconsistent answers to dissatisfaction items. As a result, FMM was unable to accurately estimate the correlation between satisfaction and dissatisfaction for this group, as shown by the large standard error in the correlation estimate. The implication is that when researchers assume population homogeneity without testing for it, inconsistent responses may lower the correlation between satisfaction and dissatisfaction, thus exaggerating the distinction between the two. We suspect that similar results may have been obtained for other constructs

embroiled in dimensionality debates. Therefore, we propose a procedure for testing population heterogeneity in the context of examining construct dimensionality (Figure 1), which we hope may shed light on these decades-old debates.

In addition to construct dimensionality, our procedure may contribute to the discussion regarding the psychometric distinction between positively and negatively worded items. Psychometricians often claim that negatively worded items are psychometrically inferior to positively worded ones, and thus it has been suggested that negatively worded items be excluded from surveys (Lindwall et al., 2012; Magazine et al., 1996; Schriesheim et al., 1991; Schriesheim & Eisenbach, 1995; van Sonderen et al., 2013). Our results are consistent with this interpretation: participants experienced more difficulty giving consistent answers to the dissatisfaction items than to the satisfaction items. Fortunately, however, this appeared to be true only for a minority of the respondents. For the majority, not only did they have no problem answering job dissatisfaction items, but metric equivalence was found between satisfaction and dissatisfaction items, meaning that the two types are highly comparable. Rather than prolonging the unfruitful debate on whether to include negatively worded items (Lindwall et al., 2012), our method helps break the impasse by directing future effort to understanding why some respondents (around 20% in the current study) have trouble answering negatively worded items.

Using FMM to examine construct dimensionality is similar to Litson et al.'s (2017) use of FMM to study trait-method interaction. Their approach was originally intended for a general MTMM model, and thus may not apply specifically in the case of testing construct dimensionality. As a result, there are substantial differences between the two approaches. First, there are procedural differences. Their strategy was to first confirm the best between-method measurement invariance model among all the single-class solutions, and then assume the validity of this measurement invariance model in a higher-class solution. The rationale behind their strategy was perhaps practical: the number of estimated parameters can substantially diminish when some level of measurement invariance is assumed, thus alleviating computational difficulties. However, our results show that measurement invariance in a lower-class solution does not necessarily extend to higher-class solutions. When there is no substantial theoretical

reason to assume that between-method measurement invariance holds across classes, the best strategy would appear to be to test invariance within each class.

The second difference from Litson et al.'s (2017) approach is that our approach encourages response pattern analysis within each class, whereas their general approach does not. Although simulation research (Schmitt & Stuits, 1985) has considered the results of data quality on construct dimensionality, few studies, with the notable exception of Reise et al. (2016), have studied the impact of problematic response pattern in real data, as well as the influence of response pattern on factor correlations. Reise et al. (2016) examined the data pattern of respondents who fit well in a bifactor model but not in a unidimensional model. They found that many such respondents gave implausible responses to regular- and reverse-keyed self-esteem items. The current study went beyond Reise et al. (2016) by conducting detailed item response pattern analysis. Whereas they found that some respondents have trouble answering positively worded items, negatively worded items, or both, we found that a large group of respondents have trouble answering negatively worded items in the current study. Using data provided by web surfers on an online site, Reise et al. (2016) could not consider the factor of careless responding in their study; as a result, their data could be severely degraded by careless responding.

Researchers, even psychometricians, often overlook the simple but important step of examining and possibly excluding careless respondents. A possible motive is researchers' reluctance to analyze smaller datasets with less power to find an effect. However, as the literature shows, careless responding can severely bias construct correlations (Kam & Meyer, 2015b; Huang et al., 2012). In a construct dimensionality study, a dataset with a substantial number of careless respondents may show attenuated relationships between positively and negatively worded items, causing a unidimensional construct to appear two-dimensional (Kam & Meyer, 2015b). After excluding careless respondents, the current study showed that *even some careful respondents*, who are generally able to give consistent answers to the satisfaction items, do not give consistent responses to the dissatisfaction items. To our knowledge, this is a novel finding.

Including careless respondents can also make the results difficult to interpret. For example, is an aberrant response pattern (e.g., 11434 among job dissatisfaction items) caused by participants' genuinely inconsistent response style or by inattentive responding? (Reise et al., 2016). To exclude careless respondents, we recommend the use of instructed response items (e.g., 'Please answer *Strongly Agree* to this item), which are simple but effective (Kam & Meyer, 2015b; Meade & Craig, 2012). When an a priori method to identify careless responding is not possible, various post-hoc methods (e.g., Mehalanobis distance combined with consecutive repeated responses) can be used (Kam & Meyer, 2015b; Meade & Craig, 2012).

As a helpful reminder, to investigate response incongruity, raw data responses instead of parceled data (as in Litson et al., 2017) should be used. Litson et al. were mainly concerned with method correlations at different levels of trait scores (i.e., trait-method interactions), whereas our purpose was in response incongruity among raw items. The use of parceled data would tend to obscure inconsistency among items.

There is a technical issue in using FMM to study population heterogeneity in construct dimensionality. Litson et al. (2017) suggest a maximum of 100,000 start values for a single-trait model. However, their recommendation is probably based on the assumption that a between-method measurement invariance model in a single-class solution can reasonably extend to a higher-class model, and thus many fewer estimators would need to be obtained in a higher-class solution. However, we did not find this assumption tenable in our data. Further, Litson et al. assigned a large number of items to three manageable parcels for their FMM analysis; this procedure drastically reduces the number of parameters to be estimated and may produce indicators with more amenable distributional properties. When item-level (as opposed to parcel-level) parameters are allowed to be freely estimated both between methods and between classes, we suspect that the number of iterations will need to increase substantially. Without systematic study, our experience suggests that 200,000 to 400,000 start values may be required when the number of parameter estimates is large.

Limitations and Future Research

Despite the substantial potential of the proposed procedure to test construct dimensionality, there are certain limitations. First, FMM is a probabilistic model that does not confirm the membership of each respondent to a particular class. When respondents' class membership was extracted for further bifactor CFA analysis, parameter estimates based on a standard CFA model will inevitably differ to some degree from those estimates given by a probabilistic model. Therefore, it is important to compare parameter estimates between an FMM and a CFA model to ensure their comparability. The present results showed that the estimates were highly similar. Nevertheless, it is currently impossible to conduct bifactor analysis while retaining the probabilistic nature of an FMM analysis: after extracting respondents' membership, their class is treated as confirmed rather than probabilistic. To the best of our knowledge, at this time there are no other methods to circumvent this issue, but we hope that future researchers can find a solution.

Second, it must be acknowledged that the present procedure is computationally intensive, involving many model comparisons and much data fitting. Each FMM model requires many start values to ensure that the best likelihood is replicable. Previous research on FMM simulations tends to fit a much simpler model and requires fewer start values. Litson et al. (2017) fit a large number of FMM models with fewer start values, but the number of parameter estimates in their investigation was smaller than in the current study. Therefore, previous studies provide little information on the optimal number of start values needed for arriving at the optimal solution. One would expect a larger number of start values when the model becomes more complex and when the number of parameter estimates (e.g., free factor loadings and residual variances) increases. We recommend future simulation research to investigate this issue.

Third, although the current study has discovered individual differences in answering positively and negatively worded items, it has not examined which individual difference variables are involved. As reviewed earlier, several potential predictors, including personality traits, cognitive ability, reading ability, and trait anxiety have been identified (Corwyn, 2000; de Jonge & Slaets, 2005; Gnambs, & Schroeders, 2017; Michaelides et al., 2015; Marsh, 1996; Quilty et al., 2006; Roszkowski & Soven, 2010; Tomás et al., 2013). Future researchers could incorporate such candidate variables in FMM analysis to

determine which ones best predict response inconsistency between positively and negatively worded items.

References

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Thousand Oaks, CA: Sage.

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods, 12*, 381–398.

Conway, J. M. (1998). Estimation and uses of the proportion of method variance for multitrait-multimethod data. *Organizational Research Methods, 1*, 209–222.

Corwyn, R. F. (2000). The factor structure of global self-esteem among adolescents and adults. *Journal of Research in Personality, 34*, 357-379.

Credé, M., Chernyshenko, O. S., Bagraim, J., & Sully, M. (2009). Contextual performance and the job satisfaction–dissatisfaction distinction: Examining artifacts and utility. *Human Performance, 22*, 246–272.

Dalbert, C., Lipkus, I. M., Sallay, H., & Goch, I. (2001). A just and an unjust world: Structure and validity of different world beliefs. *Personality and Individual Differences, 30*, 561–577.

de Jonge, P. & Slaets, J. (2005). Response sets in self-report data and their associations with personality traits. *European Journal of Psychiatry, 19*, 209–214.

Fincham, F. D., & Linfield, K. J. (1997). A new look at marital quality: Can spouses feel positive and negative about their marriage? *Journal of Family Psychology, 11*, 489–502.

Gnambs, T., & Schroeders, U. (2017). Cognitive abilities explain wording effects in the RSEE. *Assessment*.

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: do they matter? *Personality and Individual Differences, 35*, 1241–1254.

Gu, H., Wen, Z. & Fan, X. (2017). Examining and controlling for wording effect in a self-report measure: A Monte Carlo simulation study. *Structural Equation Modeling, 24*, 545–555.

Herzberg, F., Mausner, B., Peterson, R.D. and Capwell, D. F. (1957). *Job attitudes: Review of research and opinions*. Pittsburgh: Psychological Service of Pittsburgh

Hines, G. H. (1973). Cross-cultural differences in two-factor motivation theory. *Journal of Applied Psychology, 58*, 375–377.

Holden, R. R., & Fekken, G. C. (1990). Structured psychopathological test item characteristics and validity. *Psychological Assessment*, *2*, 35–40.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort respond to surveys. Journal of Business and Psychology, 27, 99–114.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*, 828–845.

Judge, T. A., Weiss, H. M., Kammeyer-Mueller, J. D., & Hulin, C. L. (2017). Job attitudes, job satisfaction, and job affect: A century of continuity and of change. *Journal of Applied Psychology, 102*, 356–374.

Kam, C. C. S. (2017). Novel insights into item keying/valence effect using latent difference (LD) modeling analysis. *Journal of Personality Assessment*. Advance online publication. doi:10.1080/00223891.2017.1369095

Kam, C., & Meyer, J. P. (2012). Do optimism and pessimism have different relationships with personality dimensions? A re-examination. *Personality and Individual Differences*, *52*, 123–127.

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

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

Kam, C., Morin, A. J., Meyer, J. P., & Topolnytsky, L. (2016). Are commitment profiles stable and predictable? A latent transition analysis. *Journal of Management*, *42*, 1462–1490.

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment, 94*, 196–204.

Litson, K., Geiser, C., Burns, G. L., & Servera, M. (2017). Examining trait × method interactions using mixture distribution multitrait–multimethod models. *Structural Equation Modeling: A Multidisciplinary Journal*, 24, 31–51.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21–39.

Magazine, S. L., Williams, L. J., & Williams, M. 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 global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*, 810-819.

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.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods, 17*, 437–455.

Michaelides, M., Koutsogiorgi, C., & Panayiotou, G. (2015). Method Effects on an Adaptation of the Rosenberg Self-Esteem Scale in Greek and the Role of Personality Traits. *Journal of Personality Assessment, 98*, 178–188.

Muthén, L.K. and Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd Ed.). CA: McGraw-Hill.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 8799–03.

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg Self-Esteem Scale method effects. *Structural Equation Modeling*, *13*, 991–17.

Raykov, T., & Marcoulides, G. A. (2016). On studying common factor dominance and approximate unidimensionality in multicomponent measuring instruments with discrete items. *Educational and Psychological Measurement*.

Raykov, T., Marcoulides, G. A., & Chang, C. (2016). Examining population heterogeneity in finite mixture settings using latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 726–730.

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research*, *51*, 818–838.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*, 137–150.

Roszkowski, M. & Soven, M. (2010). Shifting gears: consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35*, 113–130.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In *Innovations in multivariate statistical analysis* (pp. 233–247). Springer US.

Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *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., Eisenbach, R. J., & Hill, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement*, *51*, 67–78.

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and psychological measurement*, *41*, 1101–1114.

Sliter, K. A., & Zickar, M. J. (2014). An IRT examination of the psychometric functioning of negatively worded personality items. *Educational and Psychological Measurement*, *74*, 214–226.

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

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

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*, e68967.

Vautier, S., & Pohl, S. (2009). Do balanced scales assess bipolar constructs? The case of the STAI scales. *Psychological Assessment*, *21*, 187–193.

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

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, *82*, 247–252.

Table 1

Model comparison

	BIC	AIC	Entropy
Single-class			
CMI_1: Configural invariance	13300.84	13143.10	
Two-class			
CMI_2: Configural invariance	12567.84	12247.97	.953
WMI_2: Weak invariance between methods in class 1	12548.24	12245.89	.953
SMI_2: Strong invariance between methods in class 1	**12525.96**	**12241.14**	**.951**
$StMI_2$: Strict invariance between methods in class 1	improper solution		
Three-class			
CMI_3: Configural invariance	improper solution		
Structural parameters examination			
SMI_2 + Equal variances between methods in class 1	12550.39	12269.95	.952
SMI_2 + Equal means between methods in class 1	12539.90	12259.46	.957

Note. BIC = Bayesian information criterion; AIC = Akaike's information criterion; CMI = configural measurement invariance; WMI = weak (factor loading) measurement invariance; SMI = strong (factor loading + item intercept) measurement invariance; StMI = Strict (factor loading + item intercept + item residuals) measurement invariance. The best model has been bolded.

Table 2

Parameter estimates in the final solution (SMI_2)

	Majority class (*SE*) (*n* = 463 or 78.34%)		**Minority class (*SE*)** (*n* = 128 or 21.66%)	
Factor loadings				
js1	.84***	(.02)	.38*	(.16)
js2	.69***	(.04)	.86***	(.16)
js3	.84***	(.02)	.22	(.27)
js4	.86***	(.02)	.40***	(.11)
js5	.88***	(.02)	.44***	(.14)
jd1	.84***	(.03)	.25	(.18)
jd2	.72***	(.04)	.75	(.41)
jd3	.84***	(.02)	-.23	(.47)
jd4	.77***	(.02)	.11	(.17)
jd5	.84***	(.03)	.18	(.15)
Factor correlation				
js ~~ jd	.87***	(.03)	.95**	(.35)
Residual covariances (correlation)				
js1 ~~ jd1	.17*	(.08)	-.01	(.09)
js2 ~~ jd2	.32***	(.09)	.64**	(.25)
js3 ~~ jd3	.07	(.09)	.10	(.12)
js4 ~~ jd4	.003	(.08)	-.07	(.09)
js5 ~~ jd5	.12	(.08)	.04	(.10)
Item intercepts				
js1	--		--	
js2	0.59**	(0.22)	-1.30	(3.28)
js3	0.18	(0.12)	-0.39	(3.54)
js4	0.10	(0.14)	-1.37	(1.59)

js5	0.40**	(0.16)	-1.33	(1.54)
jd1	--		--	
jd2	0.59**	(0.22)	-4.10	(7.45)
jd3	0.18	(0.12)	6.44	(6.56)
jd4	0.10	(0.14)	4.68	(2.26)
jd5	0.40**	(0.16)	0.82	(2.27)
Factor means				
Js	3.91**	(0.04)	4.11***	(0.04)
Jd	4.03**	(0.05)	3.36***	(0.13)
Factor variances				
Js	0.56**	(0.07)	0.03	(0.02)
Jd	0.79**	(0.07)	0.11	(0.16)

Note. Results for item intercepts, factor means and factor variances are unstandardized estimates, while results for factor loadings, residual covariance, and factor correlation were standardized estimates.

Table 3

Response pattern analysis

Response patterns	**Sample Responses**	**% respondents in the minority class (frequency)**	**% respondents in the majority class (frequency)**
Consistent response for both js and jd items ($i_{js} = 0$; $i_{jd} = 0$)	54544 32323 55554 11222	4.69% (6)	81.20% (376)
Inconsistent response only for js items ($i_{js} = 1$; $i_{jd} = 0$)	32444 22222 23334 32322	0% (0)	3.67% (17)
Inconsistent response only for jd items ($i_{js} = 0$; $i_{jd} = 1$)	54544 12224 44444 44411	87.50% (112)	9.94% (46)
Inconsistent response for both js and jd items ($i_{js} = 1$; $i_{jd} = 1$)	44144 14114 53254 14322	7.81% (10)	5.18% (24)

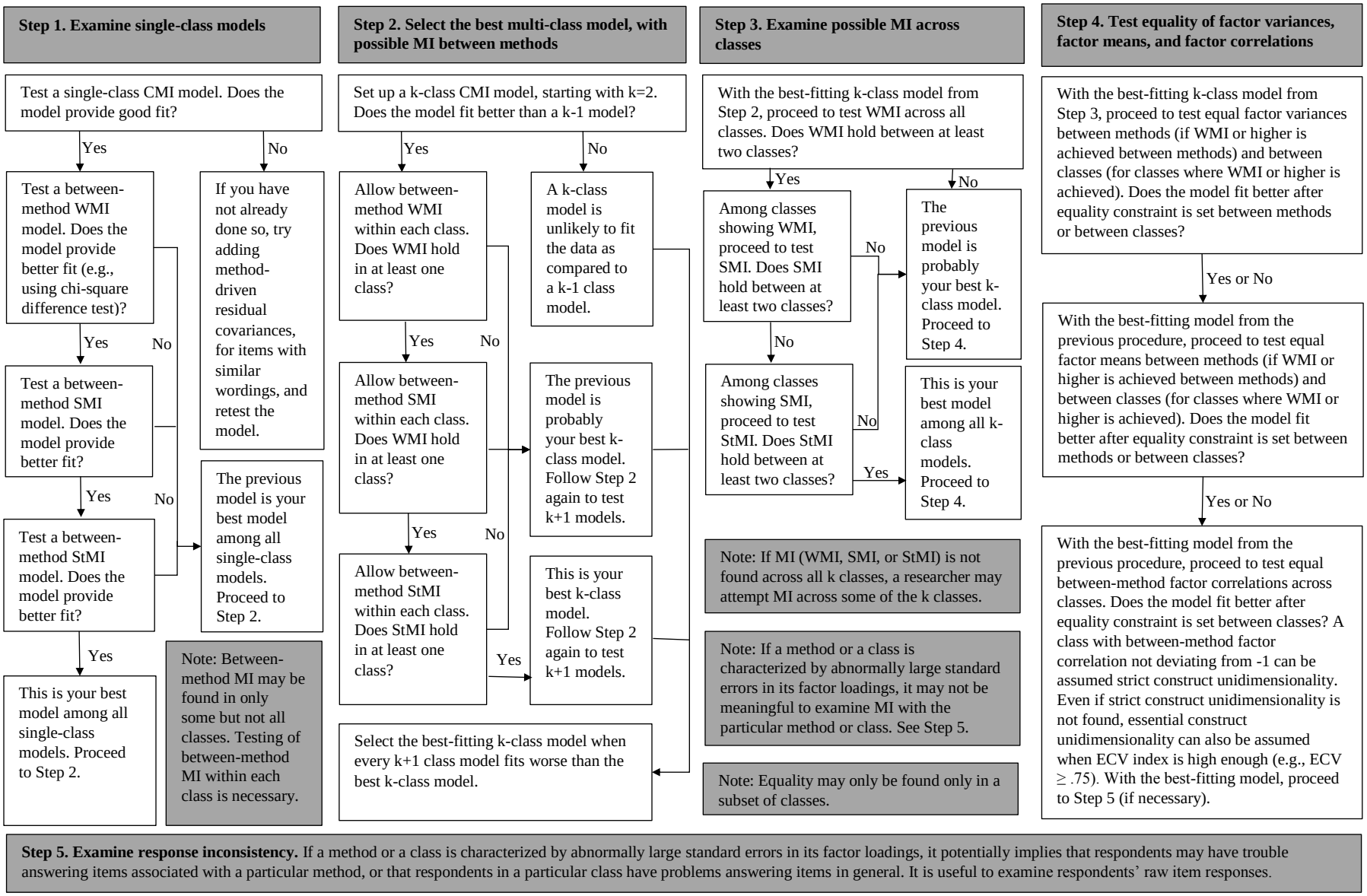

Figure 1. Flow chart for examining population heterogeneity in participants' responses to positively and negatively worded items. MI = measurement invariance; CMI = configural measurement invariance; WMI = weak (factor loading) measurement invariance; SMI = strong (factor loading + item intercept) measurement invariance; StMI = Strict (factor loading + item intercept + item residuals) measurement invariance; ECV = explained common variance.

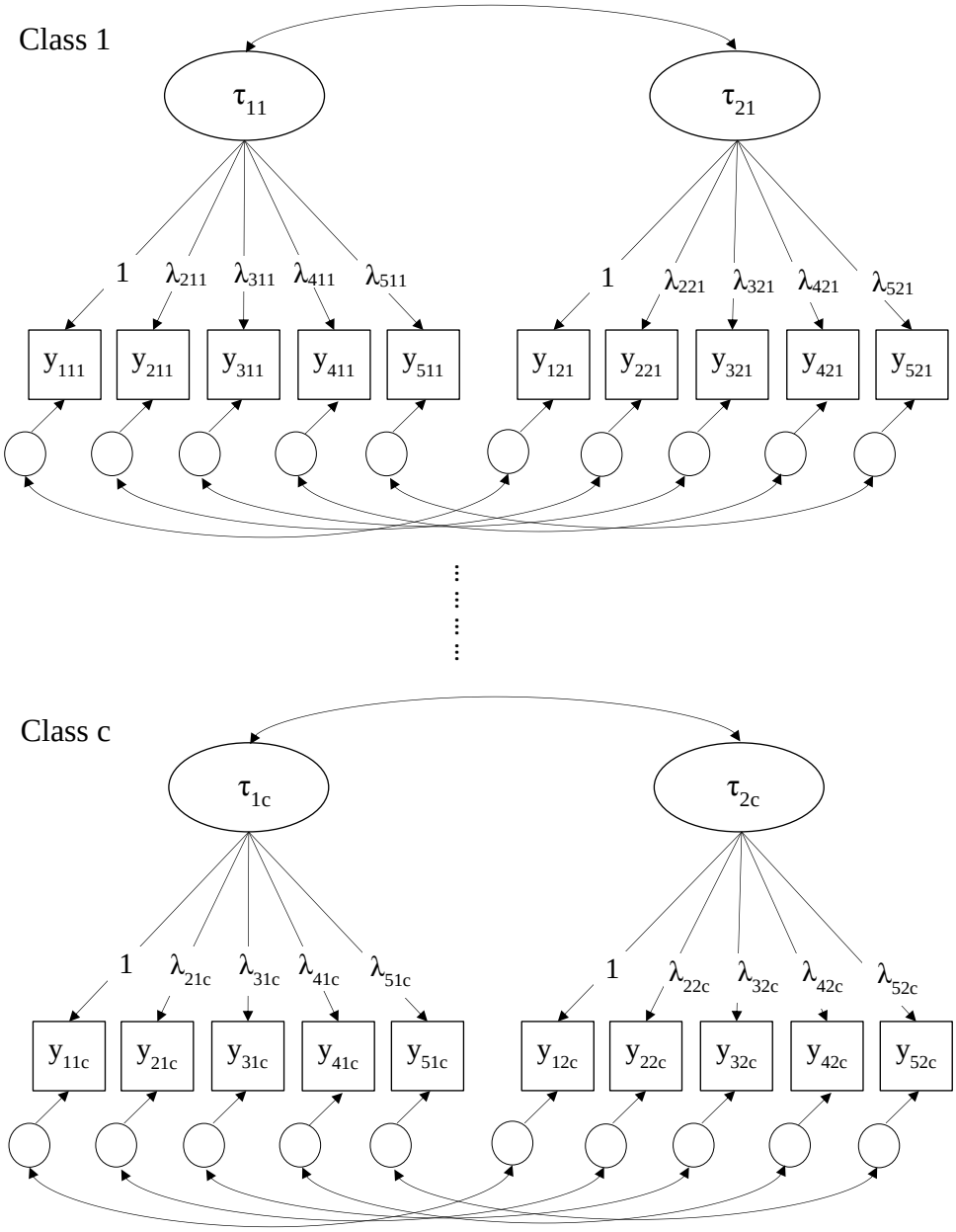

Figure 2. Factor mixture modeling for a c-class model in the current study. The correlation between methods can differ across classes. c = the c^{th} class; τ_1 = true score measured by job satisfaction items; τ_2 = true score measured by job dissatisfaction items. Circles represent residual variances. Items with parallel wording are allowed to covary with each other with the specification of residual covariances.

SUPPLEMENTARY MATERIAL S1

Mplus Syntax for Two-Class Configural Invariance Model

```
TITLE: CMI2 model
! Specifying names of your datafile here
DATA: FILE = "data.dat";

VARIABLE:
! Specifying names of the variables in your datafile
NAMES = id js1 js2 js3 js4 js5 jd1 jd2 jd3 jd4 jd5;

! Specifying the number of classes here
! with the class variable named 'c'
CLASSES = c(2);

! Specifying names of the variables in your analysis
USEVARIABLE =
js1 js2 js3 js4 js5
jd1 jd2 jd3 jd4 jd5;

! Specifying a dot as the coding for missing data
MISSING=.;

ANALYSIS:
! Specifying mixture model as the type of analytic model.
! By default, the estimator is robust maximum likelihood
TYPE = mixture;

! Specifying the initial number of random sets of starting values and
! the number of final sets for optimization.
! A large number of start value sets could avoid the problem of
! non-optimal solution (i.e., local maxima).
STARTS = 200000 20000;

! Specifying the maximum of iterations at the initial stage
! A large number of initial iterations may help to avoid the problem of
! non-optimal solution.
STITERATIONS = 200;

MODEL:

! Overall model specifications
! Input general model specifications that are applicable to all classes here
%OVERALL%

! Factor loading specifications
js by js1@1
js2 js3 js4 js5;

jd by jd1@1
```

```
jd2 jd3 jd4 jd5;

! Factor variance specifications
js jd;

! Item residual variance specifications
js1 js2 js3 js4 js5;
jd1 jd2 jd3 jd4 jd5;

! Item intercept specifications
[js1@0]
[js2 js3 js4 js5];
[jd1@0]
[jd2 jd3 jd4 jd5];

! Item residual covariance specifications
js1 WITH jd1;
js2 WITH jd2;
js3 WITH jd3;
js4 WITH jd4;
js5 WITH jd5;

! Factor covariance specification
js with jd;

! Factor mean specifications
[js jd];


! Specifications in Class 1
%c#1%

! Factor loadings freely estimated between methods,
! with the factor loading of the reference items fixed as 1 for model identification
js by js1@1
js2 js3 js4 js5 (a2-a5);

jd by jd1@1
jd2 jd3 jd4 jd5 (a7-a10);! Notice that the labels are different between job
! satisfaction items (a2-a5) and job dissatisfaction items (a7-a10). When the labels differ,
! Mplus does not constrain factor loadings to be equal between satisfaction items and
! dissatisfaction items. When a researcher wants to constrain factor loadings to
! be equal between satisfaction and dissatisfaction items, (as in the case of
! SM2/weak invariance between methods in class 1), change 'a7-a10' to 'a2-a5' here.

! Factor variance freely estimated between methods
js jd (a11-a12);

! Item residual variance freely estimated between methods
js1 js2 js3 js4 js5 (a21-a25);
jd1 jd2 jd3 jd4 jd5 (a26-a30);     ! Also notice the labels are different between
```

! satisfaction items and dissatisfaction items.

! Item intercepts freely estimated between methods, with item intercepts of the reference
! items fixed as 0, so that factor means can be freely estimated.
[js1@0]
[js2 js3 js4 js5] (a32-a35);
[jd1@0]
[jd2 jd3 jd4 jd5] (a37-a40); **! Also notice the labels are different between satisfaction items**
! and dissatisfaction items (except for the reference items). When a researcher wants to
! constrain item intercepts to be equal between satisfaction and dissatisfaction items,
! change 'a37-a40' to 'a32-a35' here.

! Item residual covariance freely estimated
js1 WITH jd1 (a41);
js2 WITH jd2 (a42);
js3 WITH jd3 (a43);
js4 WITH jd4 (a44);
js5 WITH jd5 (a45);

! Factor covariance freely estimated
js with jd (a51);

! Factor means freely estimated between methods
[js jd] (a61-a62);

! Specifications in Class 2
! Notice that all labels differ between class 1 (starting with 'a' such as 'a2')
! and class 2 (starting with 'b' such as 'b2'). When labels differ between two parameters,
! Mplus does not constrain two parameter estimates to be identical.
%c#2%

! Factor loadings freely estimated between methods,
! with the factor loading of the reference items fixed as 1 for model identification
js by js1@1
js2 js3 js4 js5 (b2-b5);

jd by jd1@1
jd2 jd3 jd4 jd5 (b7-b10); **! Notice that the labels are different between**
! job satisfaction items (b2-b5) and job dissatisfaction items (b7-b10). When the labels
! differ, Mplus does not constrain factor loadings to be equal between satisfaction items and
! dissatisfaction items.

! Factor variance freely estimated between methods
js jd (b11-b12);

! Item residual variance freely estimated between methods
js1 js2 js3 js4 js5 (b21-b25);
jd1 jd2 jd3 jd4 jd5 (b26-b30); **! Also notice the labels are different**
! between satisfaction items and dissatisfaction items.

! Item intercepts freely estimated between methods, with item intercepts of the reference

! items fixed as 0, so that factor means can be freely estimated.
[js1@0]
[js2 js3 js4 js5] (b32-b35);
[jd1@0]
[jd2 jd3 jd4 jd5] (b37-b40); **! Also notice the labels are different between**
! satisfaction items and dissatisfaction items (except for the reference items which have
! no label).

! Item residual covariance freely estimated
js1 WITH jd1 (b41);
js2 WITH jd2 (b42);
js3 WITH jd3 (b43);
js4 WITH jd4 (b44);
js5 WITH jd5 (b45);

! Factor covariance freely estimated
js with jd (b51);

! Factor means freely estimated between methods
[js jd] (b61-b62);

! Request standardized estimates and confidence interval in output
OUTPUT:
standardized cinterval svalues;

##

Note: To ensure that the weak invariance model (i.e., equal factor loadings) is based on the prior model (i.e., the 2-class configual invariance model), copy those starting values from the output of the configural invariance model and use them in the weak invariance model.

For example, the following syntax is part of the starting value output in the configural invariance model:

! Numbers below are only for illustrative purpose

%OVERALL%

<syntax skipped>

%C#1%

js BY js1@1;
js BY js2*0.92 (a2);
js BY js3*0.80 (a3);
js BY js4*0.82 (a4);
js BY js5*0.78 (a5);
jd BY jd1@1;
jd BY jd2*0.89 (a7);
jd BY jd3*0.84 (a8);
jd BY jd4*0.85 (a9);
jd BY jd5*0.80 (a10);

<syntax skipped>

! Copy the entire syntax output, change ‘a7’, ‘a8’, ‘a9’ and ‘a10’ to ‘a2’, ‘a3’, ‘a4’ and ‘a5’, and then
! use the modified syntax for weak invariance model.

! Do the same for the strong invariance and strict invariance model.

SUPPLEMENTARY MATERIAL S2

Overall Job Satisfaction items

js1. Considering everything, I am satisfied with my job.

js2. To me, my job is meaningful overall.

js3. All things considered, I consider my job to be pleasant.

js4. Overall, I like my job.

js5. On the whole, my job is good.

jd1. Considering everything, I am dissatisfied with my job.

jd2. To me, my job is meaningless overall.

jd3. All things considered, I consider my job to be unpleasant.

jd4. Overall, I dislike my job.

jd5. On the whole, my job is bad.

Note. js = job satisfaction items; jd = job dissatisfaction items.