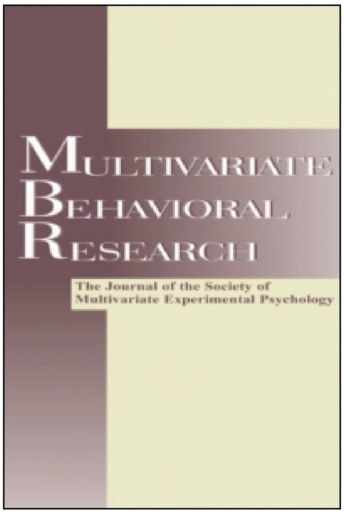

Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653673

Detecting Social Desirability Bias Using Factor Mixture Models

Walter L. Leite[a]; Lou Ann Cooper[b]
[a] Research and Evaluation Methodology Program, College of Education, University of Florida, [b] Office of Medical Education, College of Medicine, University of Florida,

Online publication date: 14 June 2010

To cite this Article Leite, Walter L. and Cooper, Lou Ann(2010) 'Detecting Social Desirability Bias Using Factor Mixture Models', Multivariate Behavioral Research, 45: 2, 271 — 293
To link to this Article: DOI: 10.1080/00273171003680245
URL: http://dx.doi.org/10.1080/00273171003680245

Multivariate Behavioral Research, 45:271–293, 2010

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171003680245

Psychology Press
Taylor & Francis Group

Detecting Social Desirability Bias Using Factor Mixture Models

Walter L. Leite
Research and Evaluation Methodology Program, College of Education
University of Florida

Lou Ann Cooper
Office of Medical Education, College of Medicine
University of Florida

Based on the conceptualization that social desirable bias (SDB) is a discrete event resulting from an interaction between a scale's items, the testing situation, and the respondent's latent trait on a social desirability factor, we present a method that makes use of factor mixture models to identify which examinees are most likely to provide biased responses, which items elicit the most socially desirable responses, and which external variables predict SDB. Problems associated with the common use of correlation coefficients based on scales' total scores to diagnose SDB and partial correlations to correct for SDB are discussed. The method is demonstrated with an analysis of SDB in the Attitude toward Interprofessional Service-Learning scale with a sample of students from health-related fields.

Researchers who develop self-report scales are frequently concerned with whether responses are affected by social desirability bias (SDB). SDB is defined as a distortion of responses in a direction considered desirable by society. This concern is evidenced by a large number of papers in the literature where a researcher's questionnaire of interest (focal scale) is administered together with

Correspondence concerning this article should be addressed to Walter L. Leite, Research and Evaluation Methodology Program, School of Human Development and Organizational Studies, College of Education, University of Florida, 1215 Norman Hall, Gainesville, FL 32611-7046. E-mail: walter.leite@coe.ufl.edu

an SDB scale in order to detect responses affected by SDB. The most commonly used SDB scales are the Marlowe-Crowne Social Desirability Scale (MCSDS; Crowne & Marlowe, 1960) and the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1991). Beretvas, Meyers, and Leite (2002) identified 1,069 published articles and dissertations since 1960 where the MCSDS or its short forms were administered in conjunction with other scales.

The most frequently used method to investigate the extent to which SDB affects the responses to a scale has been to calculate the correlation coefficient between the total SDB scores and the total scores on the focal scale (Beretvas et al., 2002). However, more sophisticated methods have been proposed to detect social desirability bias. Neil and Jackson (1970), Paulhus (1981), and Ferrando (2005) proposed methods based on principal components analysis, exploratory factor analysis (EFA), and confirmatory factor analysis (CFA), respectively. By conceptualizing a social desirability as a latent factor, these methods assume that individuals can be assigned continuous social desirability latent scores, which in turn predict the extent that their responses to a focal desirable are biased in a socially desirable direction. These methods are applicable to experimental, quasi-experimental, and observational studies.

Other methods to detect SDB either require an experimental design where a randomly assigned group of individuals is instructed to provide what they believe are socially desirable responses (Ziegler & Buehner, 2009) or require a complex manipulation of test conditions (e.g., the bogus pipeline, the randomized response technique; Thornton & Gupta, 2004). These methods have the advantage of not requiring the administration of an SDB scale but are not applicable to observational studies. The conceptualization of SDB behind these methods is somewhat different from the one behind the administration of SDB scales: SDB is considered a discrete event, where individuals may or may not change their responses in a socially desirable direction.

In this article, we propose the use of a factor mixture model (Bauer & Curran, 2004; Gagné, 2006; Lubke & Muthén, 2005, 2007; Muthén & Shedden, 1999) to detect socially desirable responding. Similarly to EFA and CFA based methods, factor mixture models have the advantage of being applicable to observational studies. The conceptualization of the socially desirable responding underlying the method we propose unifies the conceptualizations behind EFA/CFA methods and experimental/manipulation methods: individuals are associated with continuous latent scores on a social desirability trait, but whether they choose to provide a socially desirable response to an item is a discrete event. The proposed method allows the assessment of the extent that SDB presents a threat to the valid interpretation of the focal scale's scores by estimating the proportion of individuals who most likely responded in a socially desirable way in a testing situation of interest. The factor mixture model may also be used for scale development because it allows the identification of items in the scale that elicit the most socially desirable responding. Furthermore, the proposed method allows the estimation of item parameters (i.e., factor loadings, thresholds) uncontaminated by SDB. Last, the method allows for the identification of respondent characteristics or situational characteristics that predict responses that are distorted in a socially desirable direction.

SOCIAL DESIRABILITY BIAS

Although SDB has been a concern for scale developers since the 1950s (Crowne & Marlowe, 1960; Edwards, 1957), there is still no consensus in the literature about the nature of socially desirable responding, whether scales' responses should be corrected for SDB, and the best method to perform a correction. On one front, there has been a long debate in the literature (Ellingson, Sackett, & Hough, 1999; McCrae & Costa, 1983; Smith & Ellingson, 2002) about whether social desirability is a personality trait or response style. Recent studies (Holtgraves, 2004; Lönnqvist, Paunonen, Tuulio-Henriksson, Lönnqvist, & Verkasalo, 2007) as well as results of a meta-analysis (Richman, Kiesler, Weisband, & Drasgow, 1999) seem to indicate that whether an individual's response to an item is affected by SDB depends on the interaction between respondent characteristics, the nature of the scale, and the testing conditions. For example, scales with high face validity and scales about traits with negative associations (e.g., neuroticism) may be more vulnerable to SDB. On the other hand, scales with low face validity and scales about constructs not well understood by the public (e.g., locus of control) may be less affected by SDB (Furnham, 1986).

Holtgraves (2004) developed experimental studies to understand how socially desirable responding operates. In two experiments, he presented participants with items about personality traits and asked for judgments of whether those traits referred to them. He measured the response time of participants to each trait under normal, lowered social desirability and heightened social desirability conditions. The author found that individuals in the lowered social desirability condition responded faster than in the normal condition, and individuals in the heightened social desirability conditions responded slower than in the normal condition. His findings support the contention that social desirability depends on the testing situation. Holtgraves concluded that SDB is not the same as deliberate faking because respondents who "fake good" are not concerned with the veracity of their responses and therefore should not require increased response times.

The most common use of SDB scales has been to provide validity evidence in scale development. Whether it is important to control for social desirability in validity studies of a scale has also been a controversial topic. Researchers who believe SDB is a personality trait usually do not recommend controlling for SDB. On the other hand, researchers who control for SDB have applied very simple methods to do so, based mostly on classical test theory. Most frequently, researchers limit the analysis of SDB to the calculation of the correlation coefficient between the total scores on the SDB scale and the total scores, or item scores, on the focal scale (e.g., Neil & Maranda, 2007). If the correlation coefficient is not significant, researchers report it as evidence that the scores of the focal scale are not affected by SDB. When the correlation between SDB scores and the focal scale scores is statistically significant, a common correction has been to delete the scores of individuals with excessively large SDB scores (Beretvas et al., 2002). Another common correction is to compute the partial correlation between the focal scale scores and an external criterion of interest, controlling for the relationship with the scores on the SDB scale. The problem of using correlation coefficients based on scales' total scores to diagnose SDB, and partial correlations to correct for SDB, is that these methods ignore the measurement error of the scales' items and assume that items measure the construct equally well (i.e., the factor loadings/item discriminations are identical for all items). In other words, these methods assume that the items of both focal scale and the SDB scale are parallel measurements (Lord & Novick, 1968) of their respective constructs.

There have been two meta-analyses showing that controlling for SDB does not increase the predictive validity of scales' scores (Li & Bagger, 2006; Ones, Viswesvaran, & Reiss, 1996). However, the studies included in these meta-analyses are based on diagnosing SDB with correlation coefficients and controlling SDB with partial correlations between total scores, which as mentioned earlier, make strong assumptions about the psychometric properties of the scales' items. Modern psychometric methods such as confirmatory factor analysis (Bollen, 1989) and item response theory (Embretson & Reise, 2000), which account for the congeneric nature of items and their measurement error, are seldom used.

Recent research has shown changes in means, variances, and covariances due to social desirability and demonstrated the usefulness of controlling for SDB (Konstabel, Aavik, & Allik, 2006; Ziegler & Buehner, 2009). For example, Ziegler and Buehner showed that controlling for social desirability eliminates spurious mean differences between groups and correlations between constructs that should be uncorrelated. Ziegler and Buehner argued that the effects of SDB on the covariance structure are often blurred in studies of job applicants because samples of applicants to different jobs are often combined. Furthermore, Konstabel et al. argued that the failure of a number of studies to show moderator or suppressor effects for SDB scales may be due to imperfect validity of these scales rather than the lack of a social desirability effect. They were able to demonstrate that social desirability is both a suppressor and a moderator of validity and that correcting for SDB increases validity.

A limited number of latent trait methods have been proposed to detect and control for social desirability bias. Neill and Jackson (1970) proposed performing a principal components analysis on the data of the focal scale and the SDB scale. The factor with the highest loadings on the SDB items is identified as the social desirability factor and rotated to be orthogonal to the focal factor. The main limitation of Neill and Jackson's method is that a single factor with high loadings on the SDB items may not be clearly identifiable. Paulhus (1981) proposed a method for detecting SDB based on EFA that had the advantage of not depending on the administration of an SDB scale. With Paulhus's (1981) method, an SDB factor is identified among the factors extracted from the items of the focal scale (before rotation) based on the correlation between the loadings of the items and expert ratings of the extent that the items are prone to elicit socially desirable responding. The number of factors extracted should be one more than the number of factors hypothesized to account for responses to the focal scale. After the SDB factor is identified, the communality of each item is adjusted by subtracting the square of the factor loading on the SDB factor, and the solution uncontaminated by SDB can then be rotated. The main limitation of Paulhus's (1981) method is that it depends on the validity of the experts' ratings of SDB. Recently, Ferrando (2005) proposed detecting SDB with a CFA model where the items of the focal scale load on both the focal factor and the SDB factor, but the items of the SDB scale load only on the SDB factor. The cross-loadings of the focal items on the SDB factor indicate the extent to which item responses are affected by SDB.

Ziegler and Buehner (2009) pointed out that meaningful correlations between SDB scales and focal scales make it difficult to detect biasing effects of SDB. However, both Ferrando's (2005) CFA method and the method based on factor mixture models that we propose here are able to separate meaningful correlations between the SDB factor and the focal scale from the direct effects of the SDB factor on the responses to the items of the focal scale. Despite this similarity, these methods have different focus: CFA can be viewed as an item-centered diagnostic of SDB because it focuses on assessing the degree to which items are affected by SDB; factor mixture models, on the other hand, are both item centered and respondent centered because besides providing the degree of vulnerability to SDB of each item, these models also assess how likely each respondent is to provide a socially desirable response.

A FACTOR MIXTURE MODEL APPROACH FOR DETECTING SDB

Factor mixture models (B. O. Muthén & Shedden, 1999) extend the common factor model by including a latent class model that is simultaneously estimated. The latent class model posits an unobserved categorical variable that corresponds to the number of distinct populations represented in the sample data. The factor mixture model, allowing for structured means as a within-class model, was presented by Yung (1997). More complex factor mixture models have been investigated including factor mixture models conditional on covariates (Arminger, Stein, & Wittenberg, 1999; Lubke & Muthén, 2005) and mixture structural equation models in which the within-class model is a full structural equation model (Jedidi, Jagpal, & DeSarbo, 1997). Factor mixture models are recommended when a researcher hypothesizes that multiple populations generated the sample data and the population membership is unobserved. Subpopulations may differ qualitatively, such as subtypes of a medical diagnosis, or quantitatively, as in high and low scorers on a performance-based assessment designed to measure competence. Factor mixture models may include continuous or categorical outcomes and continuous or categorical covariates.

We propose the use of a factor mixture model as an extension of Ferrando's (2005) method to detect SDB. His method attempts to contrast two hypotheses: (a) a null hypothesis stating that the SDB factor is unrelated to the responses of the examinees to the focal scale and (b) an alternative hypothesis stating that the SDB factor predicts responses to the focal items for all examinees. Our method allows the examination of an intermediate hypothesis, which posits that for some individuals represented in the sample, but not all, the SDB factor predicts responses to the items of the focal scale. Considering that previous research has found that socially desirable responding depends on the interaction between respondents, the nature of the scale, and the testing conditions, the use of a factor mixture model allows conclusions about whether certain groups of examinees are more prone to respond in a socially desirable manner when responding to a focal scale administered under certain testing conditions. The fact that these conclusions do not generalize to other focal scales or testing conditions does not detract from the usefulness of the method because most test developers attempt to create standardized conditions for the administration of a scale.

With the proposed factor mixture model for detecting SDB, the null hypothesis is that respondents belong to a single class whose responses to the items of the focal scale were not affected by SDB. Several alternative hypotheses can then be evaluated: (a) the respondents belong to a single class who responded in a socially desirable way to the scale's items; (b) the respondents belong to two distinct classes, one who responded in a socially desirable way to the scale's items and one who did not; and (c) the respondents belong to three or more classes with respect to responses to the focal scale. The model associated with the null hypothesis is a single-group CFA where the focal items and the SDB items load only on their respective factors without cross-loadings. For unidimensional focal and SDB scales with categorical items, the following model represents the null hypothesis:

$$\begin{bmatrix} y^*_{i1} \\ \vdots \\ y^*_{ij} \\ y^*_{i(j+1)} \\ \vdots \\ y^*_{i(j+k)} \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \vdots \\ \tau_j \\ \tau_{j+1} \\ \vdots \\ \tau_{j+k} \end{bmatrix} + \begin{bmatrix} \lambda_{11} & 0 \\ \vdots & \vdots \\ \lambda_{j1} & 0 \\ 0 & \lambda_{(j+1)2} \\ \vdots & \vdots \\ 0 & \lambda_{(j+k)2} \end{bmatrix} \begin{bmatrix} \xi_{i1} \\ \xi_{i2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{ij} \\ \varepsilon_{i(j+1)} \\ \vdots \\ \varepsilon_{i(j+k)} \end{bmatrix}, \quad (1)$$

where $y^*_{i1} \ldots y^*_{ij}$ and $y^*_{i(j+1)} \cdots y^*_{i(j+k)}$ are the underlying continuous responses of individual i to the categorical items of the focal and SDB scales, respectively; $\tau_1 \ldots \tau_j$ and $\lambda_{11} \ldots \lambda_{j1}$ are intercepts and loadings of the j items of the focal scale; $\tau_{j+1} \ldots \tau_{j+k}$ and $\lambda_{(j+1)2} \ldots \lambda_{(j+k)2}$ are intercepts and loadings of the k items of the SDB scale; ξ_{i1} and ξ_{i2} are the individual factor scores for the focal and SDB factors, respectively; $\varepsilon_{i1} \ldots \varepsilon_{ij}$ and $\varepsilon_{i(j+1)} \ldots \varepsilon_{i(j+k)}$ are residuals.

A model that incorporates thresholds is necessary to establish the relationship between the observed categorical variable, y_{ij}, and the underlying continuous response, y^*_{ij} (Kaplan, 2000):

$$y_{ij} = \begin{cases} C_j - 1, & \text{if } \nu_{j,C_j-1} < y^*_{ij} \\ C_j - 2, & \text{if } \nu_{j,C_j-2} < y^*_{ij} \le \nu_{j,C_i-1} \\ \vdots & \\ 1, & \text{if } \nu_{j,1} < y^*_{ij} \le \nu_{j,2} \\ 0, & \text{if } y^*_{ij} \le \nu_{j,1} \end{cases}, \quad (2)$$

where C_j is the number of item categories and ν_j are threshold parameters.

The first alternative hypothesis, which states that the respondents belong to a single class who responded in a socially desirable way to the scale's items, corresponds to the model in Equation 3:

$$\begin{bmatrix} y^*_{i1} \\ \vdots \\ y^*_{ij} \\ y^*_{i(j+1)} \\ \vdots \\ y^*_{i(j+k)} \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \vdots \\ \tau_j \\ \tau_{j+1} \\ \vdots \\ \tau_{j+k} \end{bmatrix} + \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \vdots & \vdots \\ \lambda_{j1} & \lambda_{j2} \\ 0 & \lambda_{(j+1)2} \\ \vdots & \vdots \\ 0 & \lambda_{(j+k)2} \end{bmatrix} \begin{bmatrix} \xi_{i1} \\ \xi_{i2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{ij} \\ \varepsilon_{i(j+1)} \\ \vdots \\ \varepsilon_{i(j+k)} \end{bmatrix}, \quad (3)$$

where $\lambda_{12} \ldots \lambda_{j2}$ are cross-loadings of the items of the focal scale on the SDB factor, and the other terms are as defined previously.

The second alternative hypothesis is represented by a two-class factor mixture model, shown in Figure 1. With this model, the within-class means and covariance matrix are accounted for by two different measurement models. For the first class, individuals' responses to the focal scale are unaffected by SDB. This class is associated with the measurement model shown in Equation 1. For the second class, responses to the focal scale are predicted by the SDB factor. As shown in Equation 3, the measurement model for the second class specifies that all the items of the focal scale cross-load on the SDB factor. The threshold structure for both classes is specified as shown in Equation 2. To set the scale of the latent variables, it is necessary to fix either one factor loading or the factor

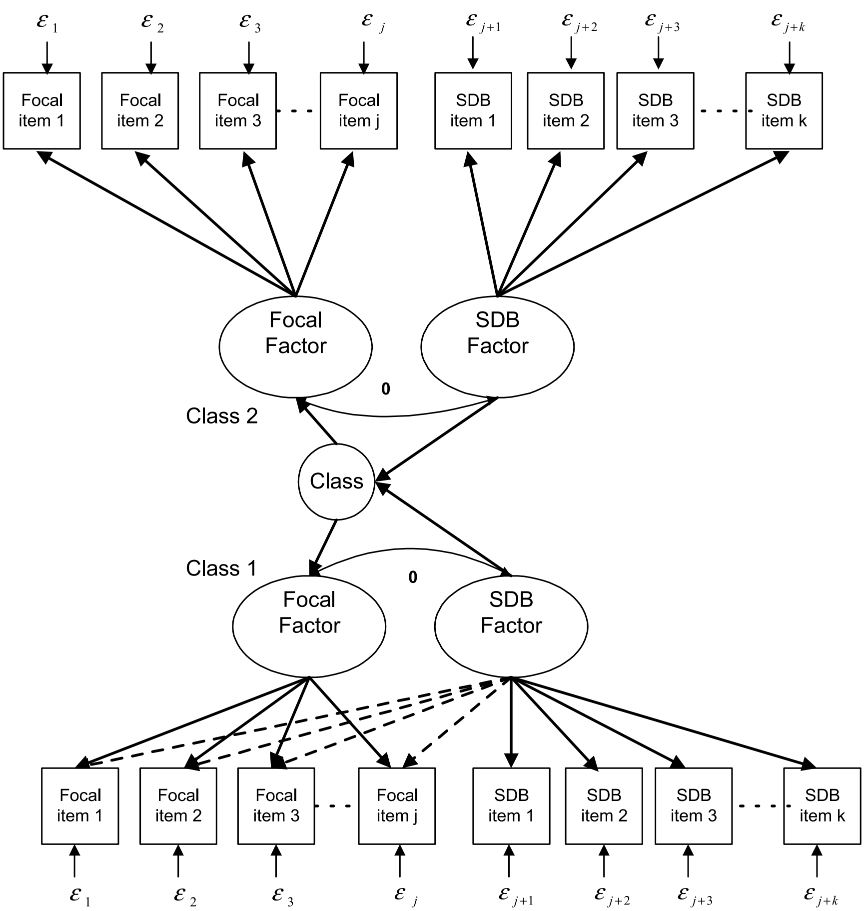

FIGURE 1 Two-class factor mixture model for detecting social desirability bias (SDB).

variance to one. It is usually more interesting to fix a factor loading to one so that the factor variance can be predicted by covariates.

Alternative hypotheses stating that individuals belong to three or more classes can also be evaluated. For the purpose of detecting SDB, it would be useful to specify a three-class model where Equation 1 describes one class and Equation 3 describes the two other classes. This analysis does not examine the presence of multiple classes that are not defined by differences in socially desirable responding. Rather, the goal of the analysis is to identify different patterns of socially desirable responding that can be captured by Equation 3.

In factor mixture models, the latent classes are unordered and modeled via a binomial variable, in the case where $m = 2$, or a multinomial variable with $M-1$ categories: $c_{im} = 1$ if participant i belongs to class m, $c_{im} = 0$ otherwise. The structural part of the factor mixture model specifies how the continuous latent variable, ξ_1, relates to the categorical latent variable, c (Lubke & Muthén, 2005; Muthén & Shedden, 1999):

$$\xi_{i1m} = \mathbf{A}_{1m}\mathbf{c}_i + \zeta_{i1m}, \tag{4}$$

where ξ_{i1m} is the score on the focal factor of individual i in class m; A_{1m} is the mean of the focal factor in class m; and ζ_{i1m} is a residual that is normally distributed with mean zero and covariance matrix, Ψ (Muthén & Shedden, 1999).

In addition, in the factor mixture model to detect socially desirable responding, the SDB factor predicts the log odds of an examinee belonging to a given class, m (i.e., the class not affected by SDB) instead of another class, M (i.e., a class responding in a socially desirable way; Lubke & Muthén, 2005):

$$\ln\left[\frac{P(c_{im} = 1|\xi_{2i})}{P(c_{iM} = 1|\xi_{2i})}\right] = \lambda_{cm} + \gamma\xi_{i2}, \tag{5}$$

where γ is a regression coefficient and λ_{cm} is an intercept that can be class specific.

The covariance between the focal factor, ξ_1, and the SDB factor, ξ_2, may be freely estimated or fixed at zero, depending on whether the researcher hypothesizes a theoretical relationship between the focal factor and the SDB factor. Freely estimating the correlation between the SDB factor and the focal factor requires an additional constraint to achieve model identification. Ferrando (2005) suggested using experts to rank the items of the focal scale with respect to the extent that they elicit socially desirable responding, then fixing the loading of the lowest ranked item on the SDB factor to zero.

Because the cross-loadings of the focal items on the SDB factor are allowed to vary between classes, allowing item parameters and the factor mean of the SDB scale to vary across classes is not necessary for the detection of socially desirable responding to the items of the focal scale. To estimate different means of the focal factor across classes, one of the classes serves as a reference and its focal factor mean is set to zero. Therefore, the estimated focal factor mean of the other classes can be interpreted as the difference in factor means between classes. Item parameters (i.e., factor loadings and thresholds) for the focal factor may be class specific or equal between classes. This allows the researcher to evaluate different levels of measurement invariance by fitting models with and without equality constraints on factor loadings and thresholds of the items of the focal scale. Factor mixture models with and without equality constraints on item parameters are compared using information criteria such as the Akaike Information Criterion (AIC), consistent AIC (CAIC), Bayesian Information Criterion (BIC), and sample size-adjusted BIC (aBIC). A comparison between models with different constraints on item parameters allows the identification of which level of measurement invariance is affected by SDB (in other words, whether SDB produces certain types of measurement noninvariance). Following the taxonomy developed by Meredith (1993), weak factorial invariance requires equal factor loadings; strong factorial invariance requires equal factor loadings and thresholds; and strict factorial invariance requires equal factor loadings, thresholds, and residual variances. Strong factorial invariance is often considered sufficient for comparing means across classes (Lubke & Muthén, 2005). Depending on the parameterization of the model (Kamata & Bauer, 2008), strict factorial invariance cannot be tested because residual variances are not freely estimated. This is the case for the delta parameterization used to fit factor mixture models with Mplus 5.1 (L. K. Muthén & Muthén, 2008).

In order to evaluate the usefulness of fitting a factor mixture model to the scores of interest, models with different numbers of classes must be compared. Because model misspecifications in a factor mixture model can be due to within-class restrictions (e.g., equality of factor loadings, thresholds, or residual variances) or to the number of classes (Lubke & Muthén, 2007), the fit of the model can be improved by either relaxing restrictions or increasing the number of classes. Although the standard likelihood ratio tests can be used to compare the fit of nested structural models, they cannot be applied to test for the optimal number of classes. Rather, the fit of models with different numbers of classes must be compared using information criteria such as the AIC, CAIC, BIC, and aBIC (Nylund, Asparouhov, & Muthén, 2007). These information criteria differ in the penalties associated with the number of parameters estimated and make different adjustments for sample size. Although different decisions may be made depending on the criteria used, a model with lower AIC, CAIC, BIC, or aBIC is preferred. The decision about the correct number of classes can also be made based on the Lo-Mendell-Rubin (LMR) test or the adjusted LMR (aLMR) test (Lo, Mendell, & Rubin, 2001). For a model estimating M classes, the LMR and aLMR estimate the probability that the data was generated by $M - 1$ classes.

A bootstrapped likelihood ratio test (BLRT) proposed by McLachlan and Peel (2000) also provides a p value that informs model choice, but this method is not available in current mixture modeling software. In a simulation study, Nylund et al. found that both the LMR and BLRT identified the correct number of classes over 80% of the time but that the BLRT consistently outperforms the LMR. They also found that although the BIC clearly outperforms all the other information criteria, the CAIC also performs adequately.

In factor mixture models, a mixing proportion, φ, estimates the proportion of the full sample that was drawn from the population of the reference class (Gagné, 2006). Details of the estimation method can be found in Muthén and Shedden (1999). The degree to which latent classes are clearly discernible, given the data and the model, can be assessed using the estimated posterior probabilities for each participant. By classifying each individual into his/her most likely class, a matrix can be constructed in which rows correspond to individuals assigned to a given class and column entries give the average conditional probabilities. Accurate classification is implied by diagonal values that are either high or low. The entropy index, a summary measure closely related to average class probabilities, provides evidence of the accuracy with which individuals have been assigned to classes. Entropy values range from zero to one with values that approach one suggesting unambiguous classifications. Lubke and Muthén (2007) provide rough guidelines for using entropy to evaluate the class assignment accuracy for a given model. Entropy values less than .60 are consistent with misclassifying 20% or more of the participants; entropy values $\geq$ .80 suggest at least 90% correct assignment.

A comprehensive evaluation of whether the interaction between items and respondents, given a certain test situation, results in socially desirable responses to the items of the focal scale can be performed by comparing the fit of multiple models with the data. Because Lubke and Muthén (2007) found that fixing factor loadings has clear beneficial effects on the estimation of factor mean differences, class proportions, correct class assignment and model convergence rates, we suggest initially fitting a two-class model that assumes strong factorial invariance for the focal factor and comparing it against the two one-class models specified according to Equations 1 and 3. If it is found that the one-class model in Equation 1 fits the data best, then there is evidence that the item responses are not affected by SDB. If the one-class model in Equation 3 has the best fit to the data, there is evidence that SDB is a general problem affecting all responses. However, the researcher may find that the two-class model provides the best fit to the data. In that case, the researcher should proceed to compare the two-class model with a three-class model and possibly models with higher numbers of classes. Once the number of classes that provides the best model fit for a model assuming strong factorial invariance is identified, it is important to compare the chosen model with models that relax invariance constraints on thresholds, loadings, and both. The examination of whether thresholds and loadings are different between classes is informative about how SDB affects the item responses.

If a decision is made to retain a multiple-class model instead of a one-class model, it is useful to attempt to predict class membership with additional covariates besides the SDB factor. Predictors of class membership allow hypothesis tests about whether certain observed groups are more likely to provide socially desirable responses to the items of the focal factor or whether certain covariates (e.g., income, age, years of experience) are related to socially desirable responding. These hypothesis tests are accomplished by adding covariates to Equation 5 and examining the statistical significance of the estimated regression coefficients. An example of the proposed method for detecting SDB is provided in the next section.

EMPIRICAL DEMONSTRATION: SDB ON RESPONSES TO THE ATTITUDES TOWARD INTERPROFESSIONAL SERVICE-LEARNING SCALE

In this example analysis, we diagnose social desirability bias on the responses to an attitudes scale used as part of the evaluation of a yearlong community-based interprofessional education course. Because the scale was administered immediately after the completion of the course, it is possible that some individuals may have felt they should respond to questions in a "community-oriented" manner that would seem to be more socially desirable.

Participants

Participants were 1st-year professional students in medicine, dentistry, pharmacy, and physical therapy and graduate students in clinical and health psychology enrolled in a community-based interprofessional education class as part of the required curriculum ($N = 668$). The Interdisciplinary Family Health (IFH) course utilizes a service-learning approach that uses the community as the context to teach core values such as community-based family health, health promotion and disease prevention, and interprofessional teamwork. Learners see firsthand the health concerns of underserved populations and must work together to effectively address real clinical problems and provide patient education (Davidson & Wadell, 2005). Community-based interprofessional learning activities are supplemented with interactive, classroom-based, small-group sessions.

Instruments

In order to use a factor mixture model to detect socially desirable responding, indicators of the SDB factor are needed. A few measures of SDB have been developed, such as the Marlowe-Crowne Social Desirability Scale (MCSDS; Crowne & Marlowe, 1960), the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1991), and the social desirability index (SDI; Hofstee, 2003). From these measures, the MCSDS has been used in thousands of studies (Beretvas et al., 2002) despite existing evidence of dimensionality problems of the full-form of the scale (Leite & Beretvas, 2005). Several short forms of the MCSDS have also been developed (Strahan & Gerbasi, 1972). From these short forms, we chose to use Strahan and Gerbasi's short form X2 (MCSDS short-X2) because it has shown promising evidence of unidimensionality in previous studies. For example, Leite and Beretvas obtained fit indices supporting the adequate fit of the unidimensional model to the MCSDS short-X2 data (i.e., CFI = .946, TLI = .939, RMSEA = .030), but the authors were cautious in providing support for the short form. A CFA of the MCSDS short-X2 using this study's data resulted in CFI = 0.905, TLI = 0.912, RMSEA = 0.054, which also support the unidimensionality of the scale, according to Marsh's recommendations for interpreting fit indices (Marsh, Hau, & Grayson, 2005; Marsh, Hau, & Wen, 2004). Previous studies have reported low reliability estimates for the scores of the MCSDS short-X2, but these estimates consisted of Cronbach's alpha, which has been shown to underestimate composite reliability if the items are congeneric (Raykov, 1997, 2001a). Therefore, we calculated reliability using CFA according to the formula provided by Raykov (2001b). The resulting reliability estimate for the scores of the MCSDS short-X2 was 0.81, which is adequate. The MCSDS short-X2 is comprised of 10 items. Five describe socially desirable behaviors and 5 describe socially undesirable behaviors. The items that describe undesirable behaviors are reverse coded.

The focal factor was measured by the Attitude toward Interprofessional Service-Learning scale (AIS). The AIS scale was developed by the Office for Interdisciplinary Education for internal use in program evaluation. Items were developed through a review of the literature on interprofessional health professions education programs in consultation with the core faculty for the course consisting of a representative of each of the five participating colleges. Items were reviewed for content relevance and representativeness with respect to the educational objectives of the course. A psychometrician evaluated individual items for clarity and students were debriefed following the initial administration to guide minor revisions. Although the course evaluation form includes additional items that vary from year to year, the responses of the students to the items that make up the AIS scale have been tracked over a 5-year period. They have been shown to be reliably sensitive to group differences in attitudes. The scale consists of 10 Likert-type items, presented as a series of statements, with five response options. The response choices represent the level of agreement with the statement and range from *strongly agree* = 5 to *strongly disagree* = 1. A representative scale item is, "As a result of this course, I feel a greater sense of responsibility to the community." The reliability of the scores, estimated using CFA, was 0.94. The AIS and MCSDS short-X2 scales were administered as part of the end-of-course evaluation.

Analysis

The Mplus 5.1 software was used for all the analyses. To identify the number of classes, we fit several factor mixture models to the data using robust maximum likelihood (MLR) estimation (L. K. Muthén & Muthén, 1998–2007). First, we fit two one-class models: (a) a one-class model without cross-loadings, described in Equation 1, and (b) a one-class model with cross-loadings (Equation 3). Then, we fit two factor mixture models assuming strong factorial invariance: (c) a two-class model according to Equations 1 and 3 and (d) a three-class model where Equation 1 describes one class and Equation 3 describes the two other classes. We also fit additional two-class models to examine measurement invariance of the AIS items across classes: (e) a model with unequal thresholds (i.e., weak factorial invariance), (f) a model with unequal factor loadings, and (g) a model with unequal thresholds and factor loadings (i.e., no measurement invariance). Because the SDB and AIS constructs are theoretically uncorrelated, all models were fit constraining the correlation to zero. To verify this assumption, the one-class models were also fit freely estimating the correlation between factors, but it was found to be nonsignificant. We compared the different models using the AIC, CAIC, BIC, aBIC, LMR, and aLMR tests. Models 1 and 2 are essentially the CFA models Ferrando (2005) proposed for assessing SDB. Therefore, the comparison of Models 1, 2, and 3 evaluated whether a factor mixture model was necessary for accounting for social desirability bias with these scores. The comparison of Models 1 to 4 allowed a decision about the number of classes that provided the best fit to the data, whereas Models 5 to 7 fine-tuned the model by investigating measurement invariance across classes.

Results

Table 1 presents information criteria and fit statistics for Models 1–7. A comparison of information criteria and p values of the LMR and aLMR test between Models 1 to 4 indicated that the two-class factor mixture model had the best fit to the data. We concluded that the data supported the hypothesis that there were two groups of respondents, one whose responses were biased in a socially desirable way and one whose responses were free of bias. Therefore, we proceeded to examine measurement invariance by fitting Models 5 to 7. We concluded that the two-class model with equal thresholds and loadings (i.e., strong factorial

TABLE 1
Fit Comparison Between Factor Mixture Models for Detecting SDB on Responses to the AIS Scale

	AIC	*CAIC*	*BIC*	*aBIC*	*LMR p Value*	*aLMR p Value*
1. One-class model without cross-loadings	21,575.14	21,960.44	21,890.44	21,668.19		
2. One-class model with cross-loadings	21,572.80	22,013.15	21,933.15	21,679.14		
3. Two-class model with equal loadings and thresholds	21,452.07	21,914.43	21,830.43	21,563.73	.03	.03
4. Three-class model with equal loadings and thresholds	21,437.18	21,971.10	21,874.10	21,566.12	.42	.43
5. Two-class model with unequal thresholds	21,291.79	21,974.32	21,850.32	21,456.61	.07	.07
6. Two-class model with unequal loadings	21,475.81	21,987.71	21,894.71	21,599.43	.00	.00
7. Two-class model with unequal loadings and thresholds	21,312.52	22,044.59	21,911.59	21,489.31	.01	.01

invariance) fit the data best. Because strong factorial invariance holds for the AIS factor, we compared the factor means between classes. We found that the mean of the AIS factor for the class most likely to respond in a socially desirable way was 0.449 standard deviations below the mean of the other class. This difference was statistically significant and can be interpreted as indicating that individuals who were most likely to respond in a socially desirable way had lower attitudes toward interprofessional service learning. The factor mixture model also allowed us to obtain factor scores for each examinee on the focal factor for each class. For this example, a pair of factor scores on the AIS factor was obtained for each student. Although we were not interested in comparisons between students, they could be performed using the score corresponding to the class the student is most likely to belong. Using factor scores to compare students in different groups is possible because the models for both classes were estimated simultaneously and the scale of the latent factors was set using the same method (i.e., by setting the loading of the same item equal to 1), which places the factor scores on the same scale of measurement.

We examined whether the factor loadings of the AIS items on the SDB factor were statistically significant. A nonsignificant loading for an item indicates that the item does not elicit socially desirable responses from the group most likely to respond in a socially desirable way. We found that all the loadings were statistically significant. Interestingly, we observed that the loadings of the first items of the AIS scale on the SDB factor were larger than the loadings of the last items. Item 2 had the highest loading (0.96) whereas Item 9 had the smallest (0.53). This could be because of the particular wording of these items and/or because students were more cautious about expressing their true thoughts with the first items but became progressively less concerned as they responded to more items (i.e., an order effect). A careful examination of the item wording provided some support for the first hypothesis because the first items refer to individual accomplishments (e.g., Item 2: I learned helpful information in home visits that I would not have obtained had I seen this family only in a health care setting.) whereas the later items refer to characteristics of the course (e.g., Item 9: The IFH course gave me the opportunity to apply knowledge and skills acquired in my other classes to real life situations.). Responding about characteristics of a course would tend to elicit less socially desirable responses than responding about individual accomplishments. The second hypothesis could be tested with another wave of data collection where the order of the items is reversed. These factor loadings, as well as the factor loadings of the AIS items on the AIS factor for both classes, are shown in Table 2.

Based on the Wald test, we found that the coefficient of the regression of class membership on the SDB factor was statistically significant ($\beta = 0.474$, $p = .002$), indicating that individuals with higher levels of the SDB factor are more likely to belong to the class that responded to the items of AIS

TABLE 2
Standardized Factor Loadings of the AIS Items on the AIS and SDB Factor for Each Class

	Class Affected by SDB		*Class Not Affected by SDB*	
Items	*AIS Factor*	*SDB Factor*	*AIS Factor*	*SDB Factor*
Item 1	0.32	0.94	0.74	0.00
Item 2	0.26	0.96	0.72	0.00
Item 3	0.57	0.78	0.64	0.00
Item 4	0.62	0.76	0.74	0.00
Item 5	0.54	0.77	0.49	0.00
Item 6	0.67	0.72	0.78	0.00
Item 7	0.72	0.64	0.71	0.00
Item 8	0.80	0.56	0.79	0.00
Item 9	0.82	0.53	0.76	0.00
Item 10	0.80	0.56	0.78	0.00

in a socially desirable way. This result is interesting for this study because it confirms that the respondents' trait levels on the SDB factor affect how they respond to the items of the focal scale. However, the estimated effect would not be expected to generalize to other studies because how strongly the SDB factor scores predict socially desirable responding depends on the testing situation, the focal scale, and the nature of the SDB trait measured by the SDB scale chosen. As mentioned in the literature review, the nature of the trait that leads to socially desirable responding is still unclear. Several interpretations of the SDB trait(s) have appeared in the literature, such as need for approval (Crowne & Marlowe, 1964), attribution and denial (Tatman, Swogger, Love, & Cook, 2009), impression management, and self-deceptive enhancement (Paulhus & Reid, 1991). The method we propose could be used to detect SDB with any set of SDB indicators, but the expected relationship between SDB factors and class membership would change depending on the nature of the indicators.

For each respondent, the posterior probabilities of belonging to each class were calculated, and individuals were assigned to a class based on highest posterior probability of membership. This classification indicated that 56 individuals in the sample (8.47%) are most likely to have provided responses confounded by SDB. The entropy of this model was 0.82, which according to Lubke and Muthén (2007) indicates that the classification was correct for at least 90% of the cases. This result indicates that socially desirable responding to the items of the AIS scale is a problem with this population of students in this testing condition (i.e., the end-of-course evaluation of the community-based interprofessional education class). Whether this problem is of practical significance depends on whether the 56 individuals most likely to respond in a socially desirable way are spread somewhat evenly across disciplines (medicine, dentistry, pharmacy, physical therapy, and psychology) or are concentrated in a specific group.

In a validation study, researchers should determine prior to data collection which external variables would be useful to include in the model to predict class membership and/or the focal factor. However, because our example analysis is not a validation study of the AIS scale, we used the available secondary institutional data on student academic discipline for a demonstration of the potential of factor mixture models to examine relationships between socially desirable responding and external variables. To identify whether the students most likely to provide socially desirable responses are associated with any particular academic discipline, we fit an additional factor mixture model. This model was a two-class model with equal thresholds and loadings that included students' discipline as dummy-coded predictors of class membership, according to Equation 5. Medicine was used as the reference category. The results indicated that the average probability of responding in a socially desirable way for medicine, dentistry, pharmacy, physical therapy, and psychology students was 0.10, 0.10, 0.05, 0.02, and 0.42, respectively. The Wald test statistics for the regression coefficients of the different disciplines indicates that psychology students were the only group in which the probability of responding in a socially desirable way was significantly different from medicine students ($p = .001$). In the observed sample, 10.1%, 10.9%, 6.9%, 1.1%, and 31.8% of medicine, dentistry, pharmacy, physical therapy, and psychology students, respectively, were classified as most likely to have responded to the items of the AIS in a socially desirable way. Based on Furnham's (1986) argument that scales with high face validity may be more affected by SDB, one explanation is that psychology students are more likely to respond in a socially desirable way because they are most familiar with attitudes assessments. As such, the AIS would have higher face validity for them.

DISCUSSION

This article presented a new method to detect social desirability bias using factor mixture models. This method has several useful characteristics: (a) it tests whether, for a given testing situation, there are multiple latent groups that differ with respect to responding to the items of a scale in a socially desirable way; (b) it identifies whether certain items are more likely to elicit socially desirable responses; (c) it allows estimation of item parameters (i.e., factor loadings, thresholds) and factor scores for examinees not confounded by SDB; and (d) it allows the investigation of whether external variables are related to socially desirable responding. The proposed use of factor mixture models can inform scale developers interested in identifying the extent that the items of a new scale elicit socially desirable responding under certain administration conditions (i.e., the focus is on the items). This method can also be used by researchers interested in detecting SDB with a population of examinees and identifying examinee characteristics that account for differences in socially desirable responding (i.e., the focus is on the respondents).

One major advantage of the method we proposed is that it separates socially desirable responding, seen as discrete events, from social desirability factor scores, seen as continuous trait levels of respondents. Furthermore, it allows socially desirable responding (i.e., class membership), the social desirability factor scores, and the focal factor scores unconfounded by SDB to be related to external variables. In assessments where there may be some contextual pressure to respond in a socially desirable way, detecting SDB with factor mixture modeling may be useful in providing validity evidence for the assessment. If the factor mixture model identifies a large proportion of individuals likely to have responded in a socially desirable way, the researcher may use the information obtained from analyzing class membership predictors to modify test instructions, confidentiality statements, item wording, and characteristics of the test site. Also, if certain items have high cross-loadings on the SDB factor, the researcher may want to modify the wording of these items. Examining cross-loadings is fundamentally different from examining the correlation between the focal factor and the SDB factor. With a factor mixture model, the focal and SDB factors may be allowed to correlate, and this is not an indication that the items elicit socially desirable responses. However, cross-loadings indicate the extent that an item elicits SDB for a given population, and the probabilities of class membership indicate how likely members of the population are to provide responses biased in a socially desirable way. Therefore, the results of an analysis using factor mixture models highlight that the common practice of diagnosing SDB by examining correlations between the SDB scale's total scores and the focal scale's total scores is inappropriate.

We have demonstrated that the factor mixture model presented in Equations 1 to 3 is able to provide useful information about socially desirable responding of examinees and vulnerability of items to SDB. This model supports research about socially desirable responding indicating that it is associated with an underlying latent trait, but it depends on the interaction between respondents, the nature of the scale's items, and the testing conditions. Because previous research has shown that there is an underlying latent trait that is measured by SDB items, the factor mixture model included an SDB factor instead of using the SDB items just as indicators of latent classes. Furthermore, because socially desirable responding occurs at the item level, factor loadings for each item of the focal scale on the SDB factor are freely estimated.

Factor mixture models are a useful tool for detecting socially desirable responding because they are able to disentangle different groups of examinees from a single sample. Although factor mixture models would seem to have wide appeal for researchers in psychology and education, they are yet to be routinely used, in part due to the unfamiliarity of applied researchers with mixture modeling. Furthermore, guidelines about what sample size is necessary to estimate a given mixture model have not been established. The necessary sample size to estimate a factor mixture model depends on several factors, such as the number of classes, number of parameter estimates, constraints, and reliability of the data (Lubke & Muthén, 2005).

Because the focus of this article was to present factor mixture models as a tool for detecting SDB and to demonstrate the method with an example, we did not perform an evaluation of its performance against external criteria for whether individuals are responding in a socially desirable way. Future research could examine whether class separation obtained with factor mixture models can be replicated with external criteria. A simulation study could also be used to investigate the power of factor mixture models to detect socially desirable responding, the sensitivity of the method to sample size, proportion of class membership, and magnitude of factor loadings. Furthermore, there are known problems with mixture models that could limit the effectiveness of the proposed method with some data. These problems include convergence problems (Gagné, 2006) and recovery of spurious latent classes when the data are nonnormal and/or there are nonlinear relationships among variables (Bauer & Curran, 2004).

A further limitation of the proposed method is that it depends on the concurrent administration of social desirability indicators together with the focal scale. A few SDB scales are available to the applied researcher (e.g., the MCSDS, Crowne & Marlowe, 1960; the BIDR, Paulhus, 1991) as well as several short forms of these scales. This study used 10 social desirability indicators, but it would be important for future research to address whether a smaller number of indicators could be used to obtain similar results. Potentially, as few as 3 indicators could work well as measures of the SDB factor.

In conclusion, this article introduced a new method based on factor mixture models for the detection of social desirability bias. SDB is one among many types of response bias, so other uses of factor mixture models for the detection of response bias are possible. This area of research is new, and several opportunities for further studies exist. Considering that social desirability bias is a common concern in scale development and validation (Beretvas et al., 2002), the proposed method has potential for widespread application.

REFERENCES

Arminger, G., Stein, P., & Wittenberg, J. (1999). Mixtures of conditional mean- and covariance-structure models. *Psychometrika*, *64,* 475–494.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods, 9,* 3–29.

Beretvas, S. N., Meyers, J. L., & Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement, 62,* 570–589.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

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

Crowne, D. P., & Marlowe, D. (1964). *The approval motive*. New York, NY: Wiley.

Davidson, R. A., & Wadell, R. (2005). An historical overview of interdisciplinary family health: A community-based interprofessional health professions course. *Academic Medicine, 80,* 334–338.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York, NY: Dryden.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84,* 155–166.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Ferrando, P. J. (2005). Factor analytic procedures for assessing social desirability in binary items. *Multivariate Behavioral Research, 40,* 331–349.

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

Gagné, P. (2006). Mean and covariance structure mixture models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 197–224). Greenwich, CT: Information Age Publishing.

Hofstee, W. K. B. (2003). Structures of personality traits. In B. Weiner, T. Millon, & M. Lerner (Eds.), *Handbook of psychology: Personality and social psychology* (Vol. 5, pp. 231–254). Hoboken, NJ: Wiley.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin, 30,* 161–172.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). STEMM: A general finite mixture structural equation model. *Journal of Classification, 14,* 23–50.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15,* 136–153.

Kaplan, D. (2000). *Structural equation modeling : Foundations and extensions*. Thousand Oaks, CA: Sage.

Konstabel, K., Aavik, T., & Allik, J. (2006). Social desirability and consensual validity of personality traits. *European Journal of Personality, 20,* 549–566.

Leite, W. L., & Beretvas, S. N. (2005). Validation of scores on the Marlowe-Crowne Social Desirability Scale and the Balanced Inventory of Desirable Responding. *Educational and Psychological Measurement, 65,* 140–154.

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

Lo, Y., Mendell, N. R., & Rubin, D. (2001). Testing the number of components in a normal mixture. *Biometrika, 88,* 767–778.

Lönnqvist, J.-E., Paunonen, S., Tuulio-Henriksson, A., Lönnqvist, J., & Verkasalo, M. (2007). Substance and style in socially desirable responding. *Journal of Personality, 75,* 291–322.

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Lubke, G., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling, 14,* 26–47.

Marsh, H. W., Hau, K., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A Festschrift for Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Erlbaum.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341.

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

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55,* 463–469.

Muthén, L., & Muthén, B. O. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Author.

Muthén, L. K., & Muthén, B. O. (2008). Mplus (Version 5.1) [Computer software]. Los Angeles, CA: Author.

Neil, C. O., & Maranda, M. (2007). The development and initial validation of the Identification of Alcohol Dependence in Women Scale. *The American Journal of Drug and Alcohol Abuse, 33,* 333–340.

Neil, J. A., & Jackson, D. N. (1970). An evaluation of item selection strategies in personality scale construction. *Educational and Psychological Measurement, 30,* 647–661.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14,* 535–569.

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

Paulhus, D. L. (1981). Control of social desirability in personality inventories: Principal factor deletion. *Journal of Research in Personality, 15,* 383–388.

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

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality & Social Psychology, 60,* 307–317.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32,* 329.

Raykov, T. (2001a). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25,* 69–76.

Raykov, T. (2001b). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology, 54,* 315–323.

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84,* 754–777.

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

Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology, 28,* 191–193.

Tatman, A. W., Swogger, M. T., Love, K., & Cook, M. D. (2009). Psychometric properties of the Marlowe-Crowne Social Desirability Scale with adult male sexual offenders. *Sexual Abuse: Journal of Research and Treatment, 21,* 21–34.

Thornton, B., & Gupta, S. (2004). Comparative validation of a partial (versus full) randomized response technique: Attempting to control for social desirability response bias to sensitive questions. *Individual Differences Research, 2,* 214–224.

Yung, Y.-F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika, 62,* 297–330.

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