Item Response Theory Models for Wording Effects in Mixed-Format Scales

Educational and Psychological

Measurement

2015, Vol. 75(1) 157–178

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414528209

epm.sagepub.com

Wen-Chung Wang¹, Hui-Fang Chen^{1,2}, and
Kuan-Yu Jin¹

Abstract

Many scales contain both positively and negatively worded items. Reverse recoding of negatively worded items might not be enough for them to function as positively worded items do. In this study, we commented on the drawbacks of existing approaches to wording effect in mixed-format scales and used bi-factor item response theory (IRT) models to test the assumption of reverse coding and evaluate the magnitude of the wording effect. The parameters of the bi-factor IRT models can be estimated with existing computer programs. Two empirical examples from the Program for International Student Assessment and the Trends in International Mathematics and Science Study were given to demonstrate the advantages of the bi-factor approach over traditional ones. It was found that the wording effect in these two data sets was substantial and that ignoring the wording effect resulted in overestimated test reliability and biased person measures.

Keywords

item response theory, wording effects, bi-factor models, Bayesian methods

Self-report inventories are commonly used to measure attitudes, interests, personalities, beliefs, and values in the social sciences. They often include both positively worded (PW) and negatively worded (NW) items to minimize acquiescence or extreme response styles (Wong, Rindfleisch, & Burroughs, 2003). Measures can be

¹The Hong Kong Institute of Education, Hong Kong SAR

²City University of Hong Kong, Hong Kong SAR

Corresponding Author:

Wen-Chung Wang, The Hong Kong Institute of Education, Lo Ping Rd., Tai Po, New Territories, Hong Kong SAR.

Email: wcwang@ied.edu.hk

influenced by wording direction, regardless of the respondent's position on the underlying latent construct (Cheung & Rensvold, 2000). When all items are PW, a respondent with a yea-saying tendency is more likely to agree with each statement than to disagree, regardless of the content, resulting in an overestimation of the true latent construct (Winkler, Kanouse, & Ware, 1982; Zuckerman, Knee, Hodgins, & Miyake, 1995). Because acquiescence is a serious threat to test validity, including both PW and NW items in an inventory is often advised to cancel out systematic response biases (DeVellis, 2005; Kieruj & Moors, 2013).

When an inventory consists of both PW and NW items that are designed to measure the same latent construct (Figure 1a), a reasonable response pattern would be: "agree" with some PW items and "disagree" with some NW items. A response pattern of agree with all items or disagree with all items are considered aberrant. Based on the assumption that PW and NW items measure the same latent construct (i.e., NW items do not contain an additional irrelevant construct), responses to NW items are simply coded reversely and then treated the same as PW items (Horan, DiStefano, & Motl, 2003; Mittag & Thompson, 2000). Respondents who strongly agree with a PW item are expected to strongly disagree with its counterpart NW item (Marsh, 1996). Unfortunately, this assumption is supported by few studies (Ory, 1982), and controversial conclusions have been drawn. While some researchers have reported that respondents are more likely to respond positively to NW items than to agree with the equivalent PW ones (Holleman, 1999; Schriesheim & Hill, 1981), others have found that respondents tend to obtain higher scores on PW items than on NW items (Weems, Onwuegbuzie, Schreiber, & Eggers, 2003) and have suggested that a respondent is more likely to agree with PW items but less likely to disagree with NW items. The influences of size and wording direction vary across individual items (Riley-Tillman, Chafouleas, Christ, Briesch, & LeBel, 2009). The unexpected effect of question polarity raises the issue of what has been measured in mixed-format items, as well as a question regarding the psychometric properties of mixed-format scales.

Experimental studies have indicated that positive and negative wordings are related to different cognitive processes in mixed-format scales. NW items require more processing time than PW items do (Clark, 1976). Mapping answers to response options in NW items is more difficult and usually takes longer (Chessa & Holleman, 2007). Participants reread the questions and response options of NW items more frequently than PW ones, leading to a medium or large effect size (Kamoen, Holleman, Mak, Sanders, & van den Bergh, 2011). Longer processing reflects processing complexity (Bassili & Scott, 1996) and demonstrates the wording effect.

Aberrant psychometric properties of NW items have been addressed in the literature. Item analysis shows that NW items demonstrate somewhat lower item-to-total correlations than their counterpart PW items (Barnette, 1999; Roszkowski & Soven, 2010) and that they subsequently reduce the internal consistency reliability (Barnette, 2000; Cronbach, 1946, 1950). Even a small proportion of NW items (e.g., 2 out of 20 items) can lower the internal consistency (Roszkowski & Soven, 2010). Exploratory and confirmatory factor analysis often support a two-factor solution against the

Figure 1. Four approaches to wording effects in mixed-format items: (a) One-factor model; (b) Two-factor model; (c) Three-factor model; (d) Bi-factor model.

unidimensionality of a measure (Reise, Morizot, & Hays, 2007; Woods, 2006), calling for further examinations on the psychometric properties of mixed-format scales.

To better describe the structure of a mixed-format scale, a two-factor solution and a bi-factor solution in factorial analysis have been proposed (Ebesutani et al., 2012; Motl & Conroy, 2000). In a two-factor solution (Figure 1b), PW items are assumed to load on one factor and NW items on the other, and thus, two subscales are differentiated (Marsh, 1996; Roszkowski & Soven, 2010; Wong et al., 2003). Unfortunately, treating PW items and NW items as measuring two distinct (but correlated) latent constructs does not match the original intent of test development (Marsh, 1988, 1996; Woods, 2006). Construct validation studies have supported a single, common underlying construct that explains responses to mixed-format scales (Marsh, 1988), and relations between the two individual factors and criteria measures were found not significantly different (Carmines & Zeller, 1979). It is not necessary to divide a mixed-format scale into two subscales. The interpretability of the “two” factors is questionable and remains unclear. A two-factor approach may not be appropriate for mixed-format scales.

One type of bi-factor representation has been applied to mixed-format scales: all items measure the same latent construct, with PW items measuring one additional specific factor and NW items measuring another additional specific factor (Lindwall, Barkoukis, Grano, Lucidi, & Raudsepp, 2012; Reise et al., 2007). Thus, three factors are involved in the test: PW items measure an overall latent construct, called θ , and a specific factor, called γ_1 ; NW items measure θ and another specific factor, called γ_2 . These three latent constructs (θ , γ_1 , and γ_2) are assumed to be mutually independent. In this approach, γ_1 is added to describe the additional effect on PW items apart from the general factor θ , and γ_2 is added to describe the additional effect on NW items apart from the general factor θ . Hereafter, this bi-factor approach is referred to as the three-factor approach because there are three latent constructs.

Several logical problems occur in this three-factor approach. First, the meaning of θ is vague. Note that there are only two types of items in such scales: PW items and NW items. Test developers often create items that are positively worded to measure a latent construct (i.e., PW items): the higher the item score, the higher the latent construct level. In consideration of response sets such as acquiescence, test developers may create items in the opposite direction (i.e., NW items): the higher the item score, the lower the latent construct level. Each item poses a directional characteristic, either positive or negative; no items are “neutral” (neither positive nor negative). For example, in the Rosenberg Self-esteem Scale (Rosenberg, 1965), the item “I feel that I am a person of worth, at least on an equal plane with others” is considered a PW item, whereas the item “At times I think I am no good at all” is considered an NW item. In the Geriatric Depression Scale (Yesavage & Brink, 1982), the item “Have you dropped many of your activities and interests?” is considered a PW item, whereas the item “Are you basically satisfied with your life?” is considered a NW item. PW and NW are relative concepts and they define each other. If all items in a test are PW, the distinction between PW and NW is pointless. Given the relative concept of PW versus NW, θ loses its conceptual foundation. Second, a test might

consist of one type of items exclusively. If the logic of the three-factor approach is adopted, every item is treated as measuring both θ and γ_1 . Thus, θ does not correspond to the latent construct the test intends to measure (e.g., mathematics interest), and γ_1 does not reflect the (positive) wording effect, either.

This study proposes using bi-factor IRT models to analyze mixed-format scales. The following sections start with a brief introduction to standard IRT models for polytomous items; then, bi-factor IRT models for mixed-format scales are described. Parameter estimation procedures and computer programs are outlined, and two real data sets are analyzed to illustrate the applications and implications of bi-factor IRT models.

Traditional Approaches to Mixed-Format Items

Responses to rating scale items or Likert-type items are ordinal and categorical rather than interval and continuous. IRT models have been developed specifically to account for categorical item responses. The generalized partial credit model (GPCM; Muraki, 1992) encompasses many existing models as special cases. Let P_{nij} and $P_{ni(j-1)}$ denote the probabilities of scoring j and $j-1$ ($j = 1, \dots, J_i$), respectively, on item i for person n . In the GPCM, the log-odds of the two probabilities are defined as

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = a_i\theta_n - \delta_{ij}, \quad (1)$$

where θ_n is the latent construct of person n ; δ_{ij} is the j th threshold of item i ; and a_i is the slope parameter of item i . θ is assumed to follow the standard normal distribution, $N(0, 1)$. If a_i is a constant (e.g., 1) for all items, Equation 1 becomes the partial credit model (Masters, 1982). If each item has two categories, Equation 1 becomes the two-parameter logistic model (Birnbaum, 1968). If a_i is a constant for all items and each item has two categories, Equation 1 becomes the Rasch model (Rasch, 1960).

When a test consists of exclusively the same wording format of items, Equation 1 can be fit to the data directly. When a test consists of both PW and NW items, a participant who is more likely to endorse PW items or less likely to endorse NW items is considered to hold a higher level on the latent construct (Lin, 2007). To actualize this consideration, the scores of NW items should be reversed. After the reverse recoding, standard IRT models such as the GPCM are fit. This approach is depicted in Figure 1a, in which all items (after recoding NW items) are assumed to measure the same latent construct. Often, this simple recoding practice cannot fully account for wording effects (Ory, 1982). Advanced IRT models are needed.

In the two-factor approach, shown in Figure 1b, PW items are assumed to measure one latent construct and NW items are assumed to measure another latent construct. Between-item multidimensional IRT models (Adams, Wilson, & Wu, 1997; Reckase, 2009) are fit:

$$\text{For PW items: } \log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = a_i\theta_{1n} - \delta_{ij} \quad (2)$$

$$\text{For NW items: } \log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = a_i\theta_{2n} - \delta_{ij} \quad (3)$$

where θ_1 and θ_2 denote the latent constructs measured by PW items and NW items, respectively; the others were defined previously. θ_1 and θ_2 are assumed to follow a bivariate normal distribution whose correlation is freely estimated. If the correlation between θ_1 and θ_2 is very high, they may be treated as unidimensional, and the one-factor (unidimensional) approach can be applied (Figure 1a). As mentioned above, this two-factor approach fails to take into account the intent of a single latent construct in test development, and it does not consider the wording effect explicitly.

In the three-factor approach, shown in Figure 1c, PW items are assumed to measure a latent construct, called θ , and a specific factor, called γ_1 ; and NW items (after recoding) are assumed to measure θ and another specific factor, called γ_2 . Specifically, the following equations are applied:

$$\text{For PW items: } \log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = a_{i1}\theta_n + a_{i2}\gamma_{1n} - \delta_{ij} \quad (4)$$

$$\text{For NW items: } \log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = a_{i1}\theta_n + a_{i2}\gamma_{2n} - \delta_{ij} \quad (5)$$

where a_{i1} and a_{i2} denote the slope parameters of item i for the corresponding θ and γ ; the others were defined previously. The meaning of θ is the major problem in this approach, as stated previously.

The New Bi-Factor Approach

In recognition of a single latent construct in test development and the relative wording effect of NW items versus PW items, we treated PW items as measuring θ and NW items as measuring θ and a specific factor called γ . This bi-factor approach is graphically presented in Figure 1d. Specifically, we fit the following equations to PW and NW items:

$$\text{For PW items: } \log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = a_i\theta_n - \delta_{ij} \quad (6)$$

$$\text{For NW items: } \log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = a_{i1}\theta_n + a_{i2}\gamma_n - \delta_{ij} \quad (7)$$

where θ is the latent trait that the inventory intends to measure; γ can be viewed as “wording propensity”; θ and γ , as usual, are assumed to be independently and normally distributed. It is necessary to make PW items act as a reference against which

NW items are compared. There is only one γ variable, rather than two, describing the wording effect of NW items relative to PW items. A zero value of γ suggests no wording effect; a positive value of γ suggests an “increment” wording effect by increasing the chance of endorsement; a negative value of γ suggests a “decrement” wording effect by decreasing the chance of endorsement. Across persons, the γ variance depicts the magnitude of interperson variation in the wording effect. Equations 6 and 7 refer to the bi-factor model for mixed-format scales (BiMFS). When the θ and γ variances are both constrained at 1 for model identification, the ratio of the two slopes (a_{i2}/a_{i1}) describes the magnitude of wording effect for item i . The larger the ratio, the greater is the magnitude of the item.

In the BiMFS, θ is measured by both PW and NW items, and γ is measured by NW items. It can be inferred that the total score of a test (after reverse coding for NW items) will be highly correlated with the θ estimate and that the sum score of NW items will be highly correlated with the γ estimate (and the θ estimate because NW items also contribute to θ). The following empirical examples demonstrate this inference.

An interesting issue in mixed-format scales is the effect of swapping PW and NW items. In other words, what would be influenced by treating PW items as NW and NW items as PW? When raw scores are used to represent persons' latent construct levels and the wording effect is not a concern, swapping PW and NW items makes no difference except the score interpretation is reversed, which is also true when uni-dimensional IRT models are fit. Where there is wording effect, swapping PW and NW items is no longer applicable because PW and NW are relative concepts and PW items must serve as a reference to investigate the wording effect of NW items relative to PW items. In practice, there is no confusion on classifying an item to PW or NW because the classifications are normally made by test developers and clearly documented in test manuals.

Multiple Inventories

In practice, multiple inventories may be administered, and each inventory has PW and NW items to measure a θ and a γ variables. With the BiMFS, one can analyze the inventories, one inventory at a time. Such a consecutive approach, although simple, is statistically less efficient than a joint approach in which all inventories are analyzed jointly (Cheng, Wang, & Ho, 2009; Wang, Chen, & Cheng, 2004). Take two inventories as an example, in which Inventory 1 involves θ_1 and γ_1 and Inventory 2 involves θ_2 and γ_2 . In the joint approach, these four variables are assumed to follow a multivariate normal distribution. Also, θ_1 and θ_2 are assumed to be correlated, whereas γ_1 and γ_2 can be assumed to be correlated or uncorrelated with each other and θ_1 and θ_2 . The following Example 2 illustrates the joint BiMFS approach to the wording effect.

Parameter Estimation and Software for the BiMFS

Many computer programs can be used to fit the BiMFS, including general programs such as WinBUGS (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012), SAS NLMIXED, and Stata GLLAMM (Rabe-Hesketh & Skrondal, 2012), as well as specific IRT programs such as ConQuest (Adams, Wu, & Wilson, 2005-2013), IRTPRO (Cai, Thissen, & du Toit, 2011), and BMIRT (Yao, 2003). Many programs use marginal maximum likelihood estimation for parameter estimation. In recent years, Bayesian Markov Chain Monte Carlo (MCMC) techniques have become popular for advanced IRT models. In this study, WinBUGS is used for parameter estimation and model evaluation. Two questions regarding the following empirical examples were of great interest to us: “Did PW and NW items measure the same latent construct?” and, if the answer was negative, “How large was the wording effect?” To answer the first question, we fitted the GPCM (Equation 1) and the BiMFS (Equations 6 and 7) to the data and conducted model comparisons, model-fit checking, residual analyses, and person outfit mean squares. To answer the second question, we examined the difference in test reliability estimates between the two models and evaluated the magnitude of the wording effect.

The deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) can be computed for model comparison. A lower DIC value indicates a better model fit. For model–data fit, correlations of residuals between pairs of items are computed to examine the assumption of local item independence, using local dependence index Q_3 (Yen, 1984):

$$d_{ni} = u_{ni} - E_i(\hat{\theta}_n), \quad (8)$$

$$Q_3 = \text{corr}(d_i, d_{i'}), \quad (9)$$

where u_{ni} is the observed score of person n on item i ; $E_i(\hat{\theta}_n)$ is the expected score of person n on item i with an ability estimate of $\hat{\theta}_n$; d_{ni} is the residual score of person n on item i ; $\text{corr}(d_i, d_{i'})$ is the correlation between the two residual scores of item i and item i' . The expected value of Q_3 is $-1/(I-1)$ (I = test length), and the expected sampling standard deviation is $\sqrt{1/(N-2)}$ (N = sample size). Last, the principal component analysis (PCA) is conducted on the residuals to check whether a substantial dimension exists after the variances of observed scores have been extracted by the GPCM or the BiMFS. When a model fits the data well, the first eigenvalue and the ratio of the first and second eigenvalues are expected to be close to unity (Chou & Wang, 2010).

Example 1: Students' Attitudes Toward Reading

Data were drawn from the 2009 Program for International Student Assessment (PISA) Hong Kong sample, with a total of 4,837 participants. The PISA is an international assessment that aims to understand the reading, mathematics, and science abilities of 15-year-olds around the world. Coordinated by the Organization for

Table 1. Comparisons of the GPCM and BiMFS Models in the PISA Hong Kong Sample.

	DIC	θ Reliability	Q_3		EVI	EV2
			Mean	SD		
GPCM	94,704	.89	−0.071	0.126	2.195	1.385
BiMFS	91,513	.71	−0.071	0.090	1.695	1.373

Note. EV1 and EV2 are the first and second extracted eigenvalues from the residuals.

Economic Cooperation and Development, PISA began in 2000, and the data were collected at 3-year intervals. PISA 2009 was the fourth assessment wave and involved 60 countries and 5 educational systems. Each cycle of data collection assessed one major subject and two minor subject areas; the focus area of 2009 PISA was reading achievement. PISA also collected information regarding the living environments and attitudes of students.

A subscale of students’ attitudes toward reading, shown in Appendix A, was analyzed. The subscale consisted of six PW and five NW items, all of which were rated by a 4-point scale: 1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, and 4 = *strongly agree*. NW items were reversely recoded. The GPCM and the BiMFS were fit using WinBUGS. As shown in Table 1, the DIC values of the GPCM and BiMFS were 94,704 and 91,513, respectively, suggesting that the BiMFS had a better fit.

As there were 11 items, the expected value of Q_3 was $-1/(11 - 1) = -0.100$, and the expected sampling standard deviation was $\sqrt{1/(4837 - 2)} = 0.014$. The empirical mean values of the GPCM and BiMFS were the same (−0.071), but the sampling standard deviation of the BiMFS (0.090) was slightly closer to the theoretical value than was the value of the GPCM (0.126). The first eigenvalue in the GPCM was 2.195, greater than the one in the BiMFS (1.695). In addition, the ratio of the first and second eigenvalues in the GPCM was 1.58, greater than the one in the BiMFS (1.23). In short, the BiMFS had a better fit than the GPCM, and the model–data fit of the BiMFS was satisfactory.

The test reliability estimates for the θ estimates were .89 and .71 for the GPCM and the BiMFS, respectively. The BiMFS was treated as the gold standard, as it had a better fit. The test reliability was overestimated when the wording effect was ignored under the standard approach (GPCM). Using the Spearman–Brown prediction formula, the 11-item scale would need additional 24 items to raise the test reliability from .71 to .89. In other words, approximately 2.3 times of the original test length would be needed for a new scale, and such an increment in test length showed the practical impact of ignoring the wording effect on the overestimation of test reliability. The test reliability for the γ estimate in the BiMFS was .58. Given that there were only five NW items, such a low test reliability was expected.

Figure 2a and b shows the relationship between the slope or threshold estimates obtained from the GPCM and the BiMFS, respectively. The GPCM and the BiMFS yielded very similar estimates in the slope parameters ($r = .95$) and threshold

Figure 2. Estimates of the GPCM and the BiMFS models in the PISA data: (a) Slope parameter estimates; (b) Item threshold estimates; (c) Person estimates.

parameters ($r = .98$). The range of item threshold estimates in the GPCM was from -5.59 to 3.72 ($M = -0.76$), narrower than the range in the BiMFS (-6.03 to 3.93 , $M = -0.95$). Figure 2c shows the relationship between the θ estimates yielded from the GPCM and the BiMFS. Obviously, the θ estimates were very similar, with evidence of a high correlation of .98. Nearly 1.5% of participants obtained a different estimate greater than 0.5, and 22.9% obtained estimates ranging between 0.2 and 0.5. When the participants were ranked according to person estimates, only 7 of 4,837 participants had the same rank order in the two models, and 60% of 4,837 participants had a change in rank order greater than 100. This means that when ranking is used as a diagnostic purpose for interventions, ignoring the wording effect might result in misclassification of participants into a specific group and such a misclassification can cause severe problems. Note that low reliability for the θ estimates can also cause

dramatic rank order changes when the data are resampled. The rank order changes were used here to highlight the practical consequences of fitting different models on individual participants.

Parameter estimates under the BiMFS are listed in the left panel of Table 2. Ratios of the secondary slope to the primary slope (a_{i2}/a_{i1}) describe the magnitudes of the wording effect for individual NW items. Items 1 (“I read only if I have to”) and 6 (“For me, reading is a waste of time”) had the smallest ratios, 0.47 and 0.64, respectively, suggesting that the wording effects of these two items were relatively small. Conversely, the ratios for Items 4 (“I find it hard to finish books”), 8 (“I read only to get information I need”), and 9 (“I cannot sit and read for more than a few minutes”) were approximately 1, suggesting relatively large wording effects.

Table 3 lists the correlations among five variables: the sum score of PW items, the sum score of NW items (after recoding), the total score, the θ estimate, and the γ estimate. The correlations were as high as .96 between the total score and the θ estimate, and .72 between the sum score of NW items and the γ estimate. In other words, the total score depicted the θ level very well, whereas the sum score of NW items depicted the γ level fairly well.

In the BiMFS, the PW items were treated as the reference and the wording effect of NW items (relative to the reference) was examined. Specifically, Equation 6 was fit to the PW items and Equation 7 to the NW items. One may wonder what the results might be if PW items and NW items are swapped. For demonstration, in the following analysis, we treated the NW items as a reference to investigate the “wording” effect of the PW items (relative to a reference). Specifically, Equation 6 was fit to the NW items and Equation 7 to the PW items (denoted as the swapped model hereafter). The DIC value of the swapped model was 91,117, smaller than that in the BiMFS (91,513), indicating that the swapped model had a better fit than the BiMFS. The better fit for the swapped model was due to more PW items (6) than NW items (5) in the test. In the BiMFS, a PW item had one slope parameter, and an NW item had two slope parameters, resulting in a total of 16 slope parameters. In the swapped model, a PW item had two slope parameters and an NW item had one slope parameter, resulting in a total of 17 slope parameters. In general, the more parameters in a model, the better the model–data fit, especially when sample sizes are large. Parameter estimates of the swapped model are listed in the right panel of Table 2. Ratios of the secondary slope to the primary slope ranged from 0.48 to 2.06. A comparison of the parameter estimates between the BiMFS and the swapped model revealed that choosing PW or NW items as the reference led to different conclusions on the wording effect. Fortunately, in practice the specifications of PW and NW items are so well recognized that there is little confusion (if any).

Two simulations were conducted to examine the influence of fitting different models to the data. The first simulation was used to investigate the fit of the BiMFS and the swapped model to data simulated from the GPCM (without wording effect). There were 11 items (6 PW items and 5 NW items) and 3,000 respondents. The data-generating GPCM, the BiMFS, and the swapped model were fit to the data. Results

Table 2. Estimates of item parameters of the BIMFS in the PISA Hong Kong sample.

Item	Normal coding					Swapped coding				
	Primary slope		Secondary slope		Slope ratio	Primary slope		Secondary slope		Slope ratio
	Estimate	SE	Estimate	SE		Estimate	SE	Estimate	SE	
1	1.33	0.05	.63	0.04	0.47	-2.88	0.08	—	—	—
						-0.43	0.04			
						2.45	0.08			
2	3.53	0.15	—	—	—	-6.03	0.23	1.35	0.07	0.48
						-1.53	0.08			
						3.92	0.16			
3	1.95	0.06	—	—	—	-3.96	0.12	1.99	0.09	1.19
						-0.72	0.05			
						3.33	0.10			
4	1.32	0.06	1.46	0.07	1.11	-3.82	0.16	—	—	—
						-1.93	0.08			
						1.32	0.06			
5	1.63	0.06	—	—	—	-3.11	0.10	0.90	0.05	0.72
						-0.57	0.04			
						3.20	0.09			
6	1.82	0.06	1.17	0.06	0.64	-5.14	0.22	—	—	—
						-3.69	0.10			
						0.76	0.05			
7	1.66	0.06	—	—	—	-3.59	0.12	0.81	0.04	0.64
						-0.85	0.05			
						2.10	0.07			
8	0.56	0.03	0.56	0.04	1.00	-2.33	0.08	—	—	—
						-0.60	0.04			
						2.21	0.06			

(continued)

Table 2. Continued

Item	Normal coding					Swapped coding								
	Primary slope		Secondary slope		Slope ratio	Thresholds		Primary slope		Secondary slope		Slope ratio	Thresholds	
	Estimate	SE	Estimate	SE		Estimate	SE	Estimate	SE	Estimate	SE		Estimate	SE
9	2.25	0.10	2.09	0.13	0.93	-5.85	0.27	2.52	0.10	—	—	—	-1.05	0.07
						-3.53	0.15						3.06	0.10
						1.24	0.08						4.89	0.19
10	0.92	0.03	—	—	—	-2.54	0.08	0.62	0.04	1.28	0.06	2.06	-2.76	0.09
						-0.47	0.03						0.53	0.04
						2.27	0.06						3.21	0.11
11	1.08	0.04	—	—	—	-2.34	0.07	0.83	0.04	0.96	0.05	1.16	-2.33	0.07
						-0.27	0.04						0.28	0.04
						2.14	0.06						2.57	0.08

Note. — = not applicable; Items 1, 4, 6, 8, and 9 are NW items.

Table 3. Correlations Among Sum Scores of PW and NW Items, Total Score, θ and γ Estimates in the PISA Hong Kong Sample.

	Sum score of NW items	Total score	θ estimate	γ estimate
Sum score of PW items	.59*	.91*	.96*	-.04*
Sum score of NW items		.87*	.73*	.72*
Total score			.96*	.34*
θ estimate				.14*

Note. PW = positively worded; NW = negatively worded.

* $p < .01$.

showed that the parameter estimates in the three models were very close to the true values, the estimates for the second slope parameters in the BiMFS and the swapped model were very close to zero, and the person measures in the three models perfectly correlated.

The second simulation was conducted to investigate the consequences of ignoring wording effect and swapping PW items and NW items. The data were simulated from the BiMFS, with 11 items (6 PW items and 5 NW items) and 3,000 respondents. The three models were fit. Results showed that the correlations in the person measures between the GPCM and the BiMFS and between the swapped model and the BiMFS became smaller than those with no wording effect (the first simulation). Furthermore, the parameter estimates in the GPCM and the swapped model were significantly biased, and the estimates for the second slope parameters in the swapped model were far away from zero (between 0.24 and 1.92). It is concluded that it does little harm to fit the BiMFS to the GPCM data when there is no wording effect. However, when a wording effect occurred, fitting the swapped model to BiMFS data yields biased parameter estimates and conclusions about wording effect, and ignoring wording effect by fitting the GPCM to BiMFS data yielded poor estimates for the item and person parameters as well.

Example 2: Students' Attitudes and Beliefs About Mathematics and Science

The second example was used to evaluate the performance of the BiMFS in multidimensional data. We used the Trends in International Mathematics and Science Study (TIMSS) 2011 for the demonstration. TIMSS aims to assess the mathematics and science knowledge and skills of 4th and 8th graders in participating countries and educational systems. The TIMSS 2011 involved 57 countries and 20 educational systems, and it was the fifth-wave data collection since 1995. All participating students and their teachers and principals responded to questionnaires regarding the students' attitudes and beliefs about mathematics and science, and learning and the living

Table 4. Comparisons of the GPCM and BiMFS Models in the TIMSS USA Sample.

Model	DIC	θ reliability		Q_3		EVI	EV2
		Math	Science	Mean	SD		
GPCM	367,400	.88	.87	−0.036	0.149	3.186	1.840
BiMFS	345,105	.87	.86	−0.044	0.097	2.045	1.641

Note. EV1 and EV2 are the first and second extracted eigenvalues from the residuals.

environment around them, before the tests were administered. In this study, we selected 8th graders from the United States sample, with a total of 10,349 students.

Two scales of students' attitude toward mathematics and science were analyzed: one for mathematics and the other for science. Each scale consisted of five PW and four NW self-evaluation items, shown in Appendix B. A 4-point Likert-type scale was used to indicate agreement with the following items: 1 = *agree a lot*, 2 = *agree a little*, 3 = *disagree a little*, and 4 = *disagree a lot*. The responses were reversely coded, wherein a high value indicates strong agreement and a low value represents strong disagreement. Two models were fit to the data. The first model was a between-item, two-dimensional GPCM, in which the correlation between mathematics and science was directly estimated, but the wording effect was ignored. The second model was a between-item, two-dimensional BiMFS, in which the wording effect was considered, the correlation between θ_1 (mathematics) and θ_2 (science) and the correlation between γ_1 and γ_2 were directly estimated.

As shown in Table 4, the BiMFS yielded a smaller DIC value (345,105) than the GPCM (367,400) did, suggesting that the BiMFS had a better fit. The BiMFS yielded a mean Q_3 value (−0.044) that was closer to the expected value ($−1/(19 − 1) = −0.056$) than the GPCM (−0.036) was; the BiMFS yielded the first eigenvalue (2.045) and the ratio of the first and second eigenvalues (1.246) that were smaller than those yielded by the GPCM (3.186 for the first eigenvalue and 1.732 for the ratio). In short, the BiMFS had a better fit than the GPCM and had a good fit to the data.

The test reliabilities for the θ_1 and θ_2 estimates yielded by the GPCM were slightly higher (.88 and .87 for math and science, respectively) than those of the BiMFS (.87 and .86, respectively). According to the Spearman–Brown prediction formula, the nine-item subscale would need one additional item to raise the test reliabilities from .86 to .87 or from .87 to .88. Such an increment in test length indicated the practical impact of ignoring the wording effect on the overestimation of test reliability. In addition, the correlation between the two latent constructs was slightly lower in the GPCM ($r = .15$, $SE = 0.01$) than in the BiMFS ($r = .22$, $SE = 0.01$). The GPCM appeared to underestimate the correlation between the attitudes toward math and science. The underestimated correlation in the GPCM was caused by ignoring the wording effect, which was absorbed into the θ variables such that the correlation was

Table 5. Item Parameter Estimates of the BiMFS in the TIMSS U.S. Sample.

Item	Primary slope		Secondary slope		Slope ratio
	Estimate	SE	Estimate	SE	
Math					
1	2.42	0.06	—	—	—
2	1.60	0.04	1.28	0.04	0.80
3	2.09	0.06	1.27	0.04	0.61
4	2.69	0.07	—	—	—
5	1.32	0.03	0.97	0.03	0.73
6	2.04	0.04	—	—	—
7	1.41	0.03	—	—	—
8	0.98	0.02	—	—	—
9	2.17	0.07	1.57	0.06	0.72
Science					
1	2.43	0.06	—	—	—
2	1.80	0.05	1.77	0.05	0.98
3	1.69	0.04	1.33	0.04	0.79
4	2.62	0.06	—	—	—
5	1.33	0.03	1.26	0.04	0.95
6	2.08	0.05	—	—	—
7	1.73	0.04	—	—	—
8	1.41	0.03	—	—	—
9	1.74	0.05	1.77	0.06	1.02

Note. — = not applicable; Item 2, 3, 5, 8, and 9 are NW items.

attenuated by the nuisance wording effect. The test reliabilities for the γ_1 and γ_2 estimates yielded by the BiMFS were .59 and .66, respectively. Given that there were only four NW items in the two tests, such test reliabilities were acceptable.

The GPCM and the BiMFS yielded very similar slope and difficulty estimates ($r = .95$ for slope estimates and .99 for threshold estimates). The range of threshold estimates in the GPCM was from -3.80 to 1.45 ($M = -0.87$), narrower than the range in the BiMFS (-4.19 to 1.80 , $M = -1.07$). Both models yielded almost similar θ estimates, with evidence of a high correlation of .97. Four percent of the students obtained a different estimate of math or science attitude greater than 0.5, and nearly 32% received a value ranging between 0.2 and 0.5. When students were ranked in terms of math attitude, 16 out of 10,349 students (0.15%) had the same rank order in the two models, and 44.64% had a change in rank order greater than 500. Only five students obtained the same science attitude rank order in the two models, and as with the math attitude findings, nearly 40% of the students had a change in rank order greater than 500. The wording effect did influence the classification of students when rank ordering was used as a criterion for intervention or training purposes, and it should not be ignored in data analysis.

Item parameter estimates under the BiMFS are listed in Table 5. Ratios of the secondary slope to the primary slope describe the magnitude of the wording effect for individual NW items. The math attitude scale had a greater wording effect than the science attitude scale. The NW items in the math attitude scale had ratios between 0.61 and 0.80, suggesting a moderate wording effect. In the science attitude scale, only one out of four NW items showed a moderate wording effect, with a ratio of 0.79. The remaining three items (Item 2, “Science is more difficult for me than for many of my classmates”; Item 5, “Science makes me confused and nervous”; and Item 9, “Science is harder for me than any other subject”) had ratios of approximately 1, suggesting a relative large wording effect.

Conclusion and Discussion

It is common to incorporate both PW and NW items in an inventory, and they are often designed to measure the same latent construct. Traditionally, after NW items are reversely coded, all items form a single scale. It is assumed that reverse coding will make NW items function as PW items; however, this assumption needs to be empirically examined. In this study, bi-factor IRT models were adopted not only to evaluate the assumption but also to account for the wording effect in mixed-format scales. The resulting θ estimates would be free from wording effects and become comparable across persons. The slope ratio of each NW item depicted the magnitude of the wording effect on the item.

Two empirical examples of the PISA and TIMSS, representing unidimensional and multidimensional data, are provided to demonstrate the applications and implications of the proposed bi-factor IRT approach. The results support the use of the bi-factor approach in evaluating the assumption of reverse coding and assessing the wording effect in mixed-format scales. The wording effect in these two examples is not trivial, in that ignoring the wording effect results in overestimation of test reliability and biased person measures. According to the slope ratios, several NW items exhibit a rather strong wording effect.

It might be of great interest to examine group difference in the wording effect. The literature shows that the endorsement of NW items is related to grouping variables, such as personality (DiStefano & Motl, 2006), race or ethnicity (Clarke, 2000), country or age (Lindwall et al., 2012), and reading comprehension or educational levels (Marsh, 1988). To investigate these variables on the wording effect, one can treat these variables as covariates and add them directly to the BiMFS to predict γ (as well as θ , which is referred to as latent regression (Adams et al., 1997)). In this study, gender was the only variable available so that we investigated gender differences in the wording effect on attitude toward reading, mathematics, and science using the two empirical examples. The results showed no statistically significant difference between genders.

The bi-factor IRT approach can be adopted to investigate the wording effects of double negative (NN) items as well. Double negative items consist of two forms of

negative statements in one sentence, and they require great effort in the cognitive process (Davis, Kellelt, Beail, & Turk, 2009). The presence of NN items has a greater adverse impact on the internal consistency and factor structure of a scale than NW items do (Johnson, Bristow, & Schneider, 2011). Even so, NN items are still in use in scales such as the Rosenberg Self-Esteem Scale (Rosenberg, 1965). When a test consists of PW items, NW items, and NN items, the bi-factor IRT approach can be revised as follows. For PW items, Equation 6 is used; for NW items, Equation 7 is used; and for NN items, the following equation is used:

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = a_{i1}\theta_n + a_{i3}\omega_n - \delta_{ij}, \quad (10)$$

where ω_n denotes the NN wording effect on person n ; a_{i3} is its slope parameter on NN item i ; the others were defined previously. The three random effects, θ , γ , and ω , are assumed to be independently and normally distributed. A large variance of ω suggests a large NN wording effect. Future studies can be conducted to investigate how the bi-factor approach can be adopted to evaluate different types of wording effects.

Appendix A

Attitude Toward Reading in PISA 2009.

Item	Content	Wording
1	I read only if I have to.	—
2	Reading is one of my favorite hobbies.	+
3	I like talking about books with other people.	+
4	I find it hard to finish books.	—
5	I feel happy if I receive a book as a present.	+
6	For me, reading is a waste of time.	—
7	I enjoy going to a bookstore that I need.	+
8	I read only to get information that I need.	—
9	I cannot sit still and read for more than a few minutes.	—
10	I like to express my opinions about books I have read.	+
11	I like to exchange books with my friends.	+

Note. + = positive wording; — = negative wording.

Appendix B

Self-Evaluation of Mathematics/Science Ability in TIMSS 2011.

Item	Content	Wording
1	I usually do well in mathematics/science.	+
2	Mathematics/Science is more difficult for me than for many of my classmates.	—
3	Mathematics/Science is not one of my strengths.	—
4	I learn things quickly in mathematics/science.	+
5	Mathematics/Science makes me confused and nervous.	—
6	I am good at working out difficult mathematics/science problems.	+
7	My teacher thinks I can do well in mathematics/science classes with difficult materials.	+
8	My teacher tells me I am good at mathematics/science.	+
9	Mathematics/Science is harder for me than any other subject.	—

Note. + = positive wording; — = negative wording.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The project was supported by the General Research Fund, Hong Kong Research Grants Council (No. 845111).

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22, 47-76. doi:10.3102/10769986022001047

Adams, R., Wu, M., & Wilson, M. (2005-2013). *ACER ConQuest 3.0.1*. Camberwell, Australia: Australian Council for Educational Research.

Barnette, J. J. (1999). Nonattending respondent effects on internal consistency of self-administered surveys: A Monte Carlo simulation study. *Educational and Psychological Measurement*, 59, 38-46. doi:10.1177/0013164499591003

Barnette, J. J. (2000). Effects of stem and likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, 60(3), 361-370.

Bassili, J. N., & Scott, S. B. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, 60, 390-399. doi:10.1086/297760

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test score* (pp. 397-479). Reading, MA: Addison-Wesley.

- Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling*. Lincolnwood, IL: Scientific Software.
- Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park, CA: Sage.
- Cheng, Y.-Y., Wang, W.-C., & Ho, Y.-H. (2009). Multidimensional Rasch analysis of a psychological test with multiple subtests: A statistical solution for the bandwidth-fidelity dilemma. *Educational and Psychological Measurement*, 69, 369-388. doi: 10.1177/0013164408323241
- Chessa, A. G., & Holleman, B. C. (2007). Answering attitudinal questions: Modelling the response process underlying contrastive questions. *Applied Cognitive Psychology*, 21, 203-225. doi:10.1002/acp.1337
- Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, 31, 187-212. doi:10.1177/0022022100031002003
- Chou, Y. T., & Wang, W. C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement*, 70, 717-731. doi:10.1177/0013164410379322
- Clark, H. H. (1976). *Semantics and comprehension*. The Hague, Netherlands: Mouton.
- Clarke, I. (2000). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior and Personality*, 15, 137-152.
- Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, 6, 475-494. doi:10.1177/001316444600600405
- Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, 10, 3-31. doi:10.1177/001316445001000101
- Davis, C., Kellett, S., Beail, N., & Turk, J. (2009) Utility of the Rosenberg Self-Esteem Scale. *American Journal on Intellectual and Developmental Disabilities*, 114, 172-178. doi: 0.1352/1944-7558-114.3.172
- DeVellis, R. F. (2005). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage.
- DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling*, 13, 440-464. doi:10.1207/s15328007sem1303_6
- Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, T. L., Damon, J. D., & Young, J. (2012). The importance of modeling method effects: Resolving the (uni)dimensionality of the loneliness questionnaire. *Journal of Personality Assessment*, 94, 186-195. doi: 10.1207/s15328007sem1303_6
- Holleman, B. (1999). Wording effects in survey research using meta-analysis to explain the forbid/allow asymmetry. *Journal of Quantitative Linguistics*, 6, 29-40. doi: 10.1076/jqul.6.1.29.4145
- Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling*, 10, 435-455. doi: 10.1207/S15328007SEM1003_6
- Johnson, J. M., Bristow, D. N., & Schneider, K. C. (2011). Did you not understand the question or not? An investigation of negatively worded questions in survey research. *Journal of Applied Business Research*, 20, 75-86.

- Kamoen, N., Holleman, B., Mak, P., Sanders, T., & van den Bergh, H. (2011). Agree or disagree? Cognitive processes in answering contrastive survey questions. *Discourse Processes*, 48, 355-385. doi:10.1080/0163853X.2011.578910
- Kieruj, N. D., & Moors, G. (2013). Response style behavior: Question format dependent or personal style. *Quality & Quantity*, 47, 193-211. doi:10.1007/s11135-011-9511-4
- Lin, T. H. (2007). Identifying optimal items in Quality of Life Assessment. *Quality & Quantity*, 41, 661-672. doi:10.1007/s11135-006-9017-7
- Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., & Raudsepp, L. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment*, 94, 196-204. doi:10.1080/00223891.2011.645936
- Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, FL: Chapman & Hall/CRC.
- Marsh, H. W. (1988). *The Self Description Questionnaire (SDQ): A theoretical and empirical basis for the measurement of multiple dimensions of preadolescent self-concept: A test manual and research monograph*. Antonio, TX: Psychological Corporation.
- Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifacts? *Journal of Personality and Social Psychology*, 70, 810-819. doi:10.1037/0022-3514.70.4.810
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher*, 29(4), 14-20.
- Motl, R. W., & Conroy, D. E. (2000). Validity and factorial invariance of the social physique anxiety scale. *Medicine & Science in Sports & Exercise*, 32, 1007-1017. doi:10.1097/00005768-200005000-00020
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176. doi:10.1177/014662169201600206
- Ory, J. C. (1982). Item placement and wording effects on overall ratings. *Educational and Psychological Measurement*, 42, 767-775. doi:10.1177/001316448204200307
- Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata* (3rd ed.). College Station, TX: Stata Press.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Reckase, M. D. (2009). *Multidimensional item response theory: Statistics for social and behavioral sciences*. New York, NY: Springer.
- Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16, 19-31. doi:10.1007/s11136-007-9183-7
- Riley-Tillman, T. C., Chafouleas, S. M., Christ, T., Briesch, A. M., & LeBel, T. J. (2009). The impact of item wording and behavioral specificity on the accuracy of direct behavior ratings (DBRs). *School Psychology Quarterly*, 24, 1-12. doi:10.1037/a0015248
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Roszkowski, M. J., & 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, 117-134. doi:10.1080/02602930802618344

- 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. doi:10.1177/001316448104100420
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 583-639. doi:10.1111/1467-9868.00353
- Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*, 116-136. doi:10.1037/1082-989X.9.1.116
- Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education, 28*, 587-607. doi:10.1080/0260293032000130234
- Winkler, J. D., Kanouse, D. E., & Ware, J. E. (1982). Controlling for Acquiescence Response Set in scale development. *Journal of Applied Psychology, 67*, 555-561. doi:10.1037/0021-9010.67.5.555
- Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the Material Values Scale. *Journal of Consumer Research, 30*, 72-91. doi:10.1086/374697
- Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*, 189-194. doi:10.1007/s10862-005-9004-7
- Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory* [computer software and manual]. Monterey, CA: CTB/McGraw-Hill.
- Yesavage, J. A., & Brink, T. L. (1982). Development and validation of a geriatric depression screening scale: A preliminary report. *Journal of Psychiatric Research, 17*(1), 37-49.
- Yen, W. M. (1984). Effect of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145. doi:10.1007/s10862-005-9004-7
- Zuckerman, M., Knee, C. R., Hodgins, H. S., & Miyake, K. (1995). Hypothesis confirmation: The joint effect of positive test strategy and acquiescence response set. *Journal of Personality and Social Psychology, 68*, 52-60. doi:10.1037/0022-3514.68.1.52