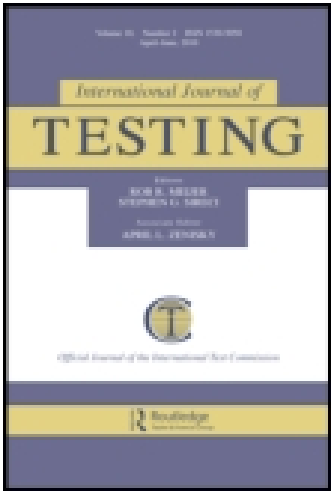

International Journal of Testing
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hijt20

A Tutorial on Interpreting Bifactor Model Scores

Christine E. DeMars [a]
[a] Center for Assessment and Research Studies, James Madison University
Published online: 06 Sep 2013.

To cite this article: Christine E. DeMars (2013) A Tutorial on Interpreting Bifactor Model Scores, International Journal of Testing, 13:4, 354-378, DOI: 10.1080/15305058.2013.799067

To link to this article: http://dx.doi.org/10.1080/15305058.2013.799067

International Journal of Testing, 13: 354–378, 2013

ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2013.799067

Routledge
Taylor & Francis Group

A Tutorial on Interpreting Bifactor Model Scores

Christine E. DeMars
Center for Assessment and Research Studies, James Madison University

This tutorial addresses possible sources of confusion in interpreting trait scores from the bifactor model. The bifactor model may be used when subscores are desired, either for formative feedback on an achievement test or for theoretically different constructs on a psychological test. The bifactor model is often chosen because it requires fewer computational resources than other models for subscores. The bifactor model yields a score on the general or primary trait measured by the test overall, as well as specific or secondary traits measured by the subscales. Interpreting the general trait score is straight-forward, but the specific traits must be interpreted as residuals relative to the general trait. Trait scores on the specific factors are contrasted with trait scores on a simple-structure model with correlated factors, using example data from one TIMSS test booklet and a civic responsibility measure. The correlated factors model was used for contrast because its scores correspond to a more intuitive interpretation of subscores, and thus it helps to illustrate how the bifactor scores should NOT be interpreted. Estimation details are covered in an appendix.

Keywords: bifactor, multidimensional IRT, subscores

The bifactor model (Gibbons & Hedeker, 1992) specifies a general factor measured by all test items as well as specific factors accounting for the residual variance shared by subsets of items. The bifactor model is increasingly being used in measurement research in education and psychology (recent examples include Armon & Shirom, 2011; Bear, Gaskins, Blank, & Chen, 2011; Betts, Pickart, & Heistad, 2011; Brown, Finney, & France, 2011; Cai, Yang, & Hansen, 2011; Kim, Sherry, Lee, & Kim, 2011; Rijmen, 2010). In the bifactor model, item responses are a function of a general or primary factor and no more than one specific or secondary factor. The specific factors are orthogonal to the general factor. The

Correspondence should be sent to Christine E. DeMars, Center for Assessment and Research Studies, MSC 6806, James Madison University, Harrisonburg, VA 22807, USA. E-mail: demarsce@jmu.edu

specific factor represents variance common to a group of items beyond the factor measured by the scale as a whole. The specific factors may be due to intended common content, such as subscales. Alternatively, they may be due to nuisance factors, such as a shared reading passage or other stimulus, or a response set such as negative wording. When the specific factors represent subscales, an alternative to the bifactor model might be a simple structure model, with each item loading on only one factor and the general factor essentially represented by the correlations among the factors. In this exposition, this model will be labeled the correlated factors model. The correlated factors model is not the focus of this demonstration, but it is included because it provides a contrast in trait score interpretation, one that is perhaps more intuitive to score users. Computationally, the bifactor model is much simpler to estimate than the correlated factors model (see Appendix), which is one reason for its increasingly common use. The purpose of this essay is to illustrate the interpretation of bifactor score estimates and how these differ from the interpretations of score estimates from a correlated factors model.

Multidimensional IRT Models

IRT models can be expressed as logistic or normal models. The logistic model is displayed here because it corresponds to the software used in the examples. For dichotomous items, the multidimensional extension of the 3-parameter logistic (3PL) model is:

$$\mathrm{P_i}(\boldsymbol{\theta}) = \mathrm{c_i} + (1 - \mathrm{c_i})\frac{\mathrm{e}^{(\mathbf{a_i'}\boldsymbol{\theta} + \mathrm{d_i})}}{1 + \mathrm{e}^{(\mathbf{a_i'}\boldsymbol{\theta} + \mathrm{d_i})}}, \tag{1}$$

where $\mathrm{P_i}(\boldsymbol{\theta})$ is the probability of correct response to item i conditional on $\boldsymbol{\theta}$, the vector of K trait scores or abilities; $\mathbf{a}_\mathrm{i}$ is a vector of K item discriminations (unstandardized loadings); d_i is the item easiness; and c_i is a lower asymptote which represents the probability of correct response for examinees with very low abilities or trait scores. The item difficulty and discrimination parameters from this model are approximately 1.7 times those from a normal model; the factor 1.7 can be used in the logistic model to place its parameters on approximately the same metric as those from a normal model. For a unidimensional model, the item difficulty $b_\mathrm{i} = -d_\mathrm{i}/a_\mathrm{i}$. In unidimensional models, with $c_\mathrm{i} = 0$, b_i corresponds to the θ at which $\mathrm{P_i}(\theta) = .5$. However, the meaning of d_i is different; with $c_\mathrm{i} = 0$, d_i is the log-odds of correct response for examinees with all elements of $\boldsymbol{\theta} = 0$. Either b_i or d_i can be used with unidimensional items, but items will generally be ordered differently by the two indices.

Figure 1 is a diagram of the bifactor model and the simple-structure correlated factors model for an instrument with three subscales. Both of these models can

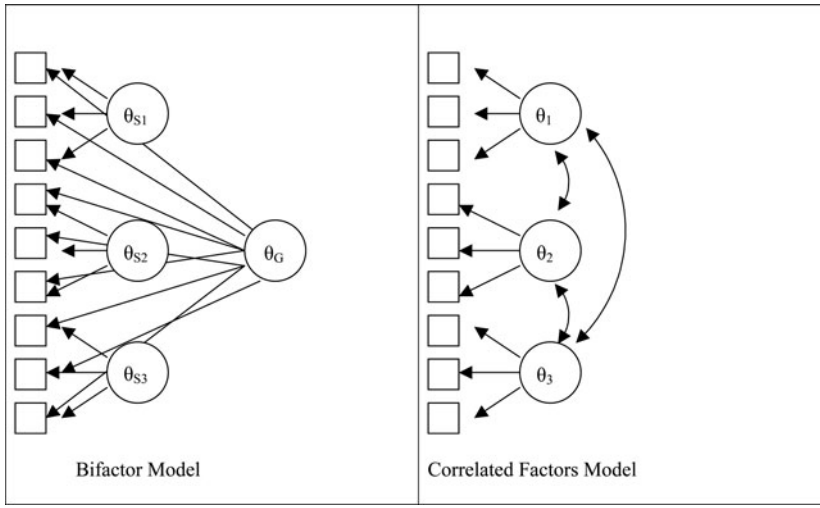

FIGURE 1
Schematics of a bifactor model and a correlated factors model. In the left panel, θ_G represents the general factor and $\theta_{S1}-\theta_{S3}$ represent the specific factors. In the right panel, $\theta_1-\theta_3$ represent the factors. The unlabeled boxes represent the observed item scores. Residual variances for the observed scores are omitted for clarity.

be represented using variants of Equation 1. Another alternative not detailed in this piece is a second-order model, which is equivalent to a bifactor model with proportionality constraints (Rijmen, 2010). The trait score estimates from the second-order model are interpreted similarly to those from the correlated factors model and will not be discussed further.

For the bifactor model, a_1 is the discrimination for the general factor. For any given item, only one of the remaining *K-1* elements of $\mathbf{a}_i$, corresponding to the specific factors, is nonzero. Using g to index the general factor and s to index whichever specific factor applies to item i, the bifactor model can be specified:

$$P_i(\theta_g, \theta_s = c_i + (1 - c_i)\frac{e^{(a_{ig}\theta_g + a_{is}\theta_s + d_i)}}{1 + e^{(a_{ig}\theta_g + a_{is}\theta_s + d_i)}}. \quad (2)$$

For the correlated factors model with simple structure, only one of the elements of $\mathbf{a}_i$ is nonzero for each item; the elements of $\boldsymbol{\theta}$ corresponding to the zero elements of $\mathbf{a}_i$ do not contribute directly to the response probability, but each trait contributes to the estimation of the other traits through their correlation. Using k to index

whichever trait applies to item *i*, the simple structure model can be specified:

$$P_i(\theta_k) = c_i + (1 - c_i)\frac{e^{(a_{ik}\theta_k + d_i)}}{1 + e^{(a_{ik}\theta_k + d_i)}}. \quad (3)$$

For polytomous items with ordered categories, the multidimensional extension of the graded response model is:

$$P_{ih}^{+}(\boldsymbol{\theta}) = \frac{e^{(\mathbf{a}_i'\boldsymbol{\theta} + d_{ih})}}{1 + e^{(\mathbf{a}_i'\boldsymbol{\theta} + d_{ih})}}, \quad (4)$$

where $P_{ih}{}^{+}(\boldsymbol{\theta})$ is the probability of choosing or scoring category *h* or higher on item *i* conditional on $\boldsymbol{\theta}$, d_{ih} is a threshold parameter for category *h* in item *i*, and the vectors **a** and $\boldsymbol{\theta}$ are as defined for Equation 1. Within each item, there is one less d_{ih} than the number of categories, because $P_{i0}{}^{+}(\boldsymbol{\theta})$ is necessarily equal to 1 for all $\boldsymbol{\theta}$. A multidimensional generalization of the partial credit model has also been developed, but the graded response model was arbitrarily chosen for this demonstration.

For both the correlated factors model and the bifactor model, the model is typically identified by constraining each factor to mean = 0 and standard deviation = 1. For the bifactor model, the specific factors are orthogonal to the general factor and to each other.

Interpreting the Bifactor Trait Score Estimates

In some contexts, the bifactor model may be chosen because the user wants to interpret the general trait scores, controlling for nuisance factors. For example, on an attitude survey, the negatively worded items might load on a specific factor, as in Kim and colleagues' (2011) study of a Korean adaptation of measures of adult attachment. For this type of use, trait score estimates on the general factor will represent more pure measures of the trait of interest after controlling for this negative-wording factor. Or on a reading test, the items related to a shared reading passage might load on a specific factor. Test users would be interested in the general reading trait score, not the specific factor scores. For example, Rijmen (2010) applied the bifactor model to an international English assessment because the items were grouped into testlets. Similarly, the Programme for International Student Assessment (PISA) tests are built using items clustered around scenarios. The scenarios are included to give real-life context, but they do not represent meaningful cognitive traits (Cai et al., 2011). Controlling these nuisance traits yields a more meaningful general factor.

In other contexts, the residual trait scores on the specific factors may be of interest. For example, on a kindergarten assessment, Betts and colleagues (2011)

assessed whether the residual parts of literacy and numeracy predicted later reading and math achievement after controlling for a general factor. In another example, Armon and Shirom (2011) posited that various personality factors would each be associated with both the general factor of work-related vigor and one of the residual factors of physical strength, emotional energy, and cognitive liveliness. Returning to the negative-wording example, researchers may be interested in examinees who score high on the negative items after controlling for the general trait, perhaps as a result of inattention or reactivity. Turning to an academic context, researchers might be interested in students who score low on geometry, after controlling for mathematics proficiency in general, to study how geometry is cognitively different from other areas of mathematics.

But in yet other contexts, score users may be interested in the whole trait measured by the subscale, not just the residual part of the score left after controlling for the general trait. Gibbons and colleagues (2007, p. 10) noted that the specific trait estimates would be underestimates of the "unconditional subdomain estimates." Warning that the scores are underestimates presumes a possibility score users will misinterpret the specific factor scores as the overall trait measured by the subscale. Returning to the mathematics example, it could be misleading to give a moderately high geometry trait score to a student with a low general score. The student might not understand that the geometry trait score was only the part of geometry independent of general mathematics proficiency. Intuitively, score users may try to use a correlated factors interpretation. The bifactor model might still be used in this context because of its computational advantages over the correlated factors model, but interpretation of the resulting trait scores would be problematic. One alternative would be to estimate subscale scores as a composite of the general trait score and the specific trait score, similar to Haberman's (2008) method of weighting the observed total and observed subscale score. Calculating the composite is computationally simpler for the bifactor model than for observed scores because the true trait scores are uncorrelated, but the meaning of the composite is similar. An alternative composite was suggested by an anonymous reviewer: the expected observed score, often called the *true score*, for the items on a subscale using both the estimated general trait score and the specific trait score. The obvious drawback to composite scores is that the resulting score estimates are far less distinct and less likely to show differential validity. But the same caveat applies to scores from the correlated factors model. If there are high correlations among the factors, the estimates of the factor trait scores lose their distinct meanings. One could reasonably argue that subscores should therefore not be reported and scores might be better modeled with a unidimensional congeneric model. However, particularly for primary and secondary school tests, subscores are increasingly in demand even when they add little measurement value (see Sinharay, 2010, or Sinharay, Puhan, & Haberman, 2011, for a discussion of this point).

Estimating a Composite Score. To form a composite score based on the trait score estimates, the weights for the general and specific trait scores can be selected based on the relative contributions of the general and specific traits to the observed responses. For an individual item, Reckase (1985, 1997; Reckase & McKinley, 1991) defined the direction of greatest slope. For the bifactor model, the angle relative to the general factor is $\alpha_i = \arccos \frac{a_{i1}}{\sqrt{a_{i1}^2 + a_{is}^2}}$, where α_i is the angle with the axis for the general factor, a_{i1} is the discrimination parameter for the general trait, and a_{is} is the discrimination parameter for specific trait s. For the subscale as a whole, the weights for the reference composite are the elements of the eigenvector of the **a**-matrix (Wang, 1985, as cited in Reckase, 2009, p. 126). The direction of the reference composite is approximately the average angle for the items loading on trait s. The composite is conceptually similar to the score on a second-order factor, but the bifactor model allows items to vary in the degree to which they load more highly on the general factor or the specific factor. If the ratio of the general factor discrimination to the specific factor discrimination were equal for all items within a specific factor, the second-order model would be equivalent to the bifactor model (Rijmen, 2010), the angle of best measurement would be the same for all items, and the composite score would be the same as the second-order factor score.

A composite score could also be estimated in the observed score metric: $\tau_s = \sum_{i \in s} P_i(\hat{\theta}_g, \hat{\theta}_s)$, where τ_s is an examinee's model-predicted observed score, often labeled the true score for historical reasons, for subscale s and P_i is the probability of correct response on item i given the examinee's score estimates on the general trait and the specific trait s. For items with more than two score categories, this generalizes to $\tau_s = \sum_{i \in s} \sum_h h P_{ih}(\hat{\theta}_g, \hat{\theta}_s)$, where h runs from 0 to 1 less than the number of score categories for item i. The τs have a somewhat nonlinear relationship with the θs, so the linear properties of the scale no longer strictly hold and the units may be compressed or expanded, particularly at the ends of the scale. However, the scale may be more interpretable to score consumers. Additionally, instead of weighting all items in the direction of the best reference composite, this method weights the general and specific factors differentially for each item in the direction of the angle of best measurement for that individual item.

Purpose

The purpose of this tutorial is to provide examples of score interpretations for bifactor model trait score estimates. To aid in understanding, the bifactor score estimates will be compared to score estimates from a unidimensional model and a correlated factors model. These contrasts will help to expose possible misconceptions about the bifactor trait scores. Similarities between the scoring methods will also be illustrated through an examination of composite scores based on

the bifactor model. Although these examples do not present any new theory or methodological developments, they may extend practitioners' understanding of the trait score estimates.

EXAMPLES

TIMSS Science, Grade 4

The TIMSS Grade 4 science test was selected to illustrate the bifactor model, using multiple choice and constructed response subscales. There is increasing demand for achievement test score reports to include subscale scores, even for tests that empirically are nearly unidimensional (Sinharay, 2010; Sinharay et al., 2011). Considering the test blueprint of the TIMSS Grade 4 science test, subscales based on the content areas of earth, life, and physical science might seem reasonable, but preliminary analyses showed the latent correlations among these subscales to be .96–1.00, which was too high for meaningful illustration of any multidimensional model so the multiple choice and constructed response scales were chosen for this example instead. Hypothetically, the context is one in which the score users have requested reports of trait score estimates for multiple choice and constructed response items, with these subscales chosen based on the needs of the end users for subscale scores. Yao (2010, p. 340) noted that it is common to report subscores based on the intended structure of the test, regardless of the empirical dimensionality of the test; that is the assumed context here. Data from the 3308 students who were assigned Booklet 3 in the 10 highest-scoring countries were used in the analysis. These countries were used because their average ability level was well matched to the item difficulties. Missing responses in the middle of the test were scored as incorrect; missing responses at the end of the session (not-reached) were treated as if the items had not been administered. The selected booklet had 25 items. Two pairs of items each referred to a common context; the items within a pair were summed and treated as a single item, resulting in a total of 23 items—11 multiple choice and 12 constructed response. Of the constructed response items, seven were worth one point and five were worth two points.

Item Parameter Estimation. Both the bifactor and correlated factor models were run in flexMIRT 1.88 (Cai, 2012). For each dimension, 49 quadrature points were used, equally spaced from –6 to 6. For the bifactor model, the priors for the slopes were set to N (1.3, .8) for the general factor and lognormal (–.4, .3) for the specific factors. For the correlated factors model, the priors for the free slopes were set to N (1.3, .8). In both models, the priors for the lower asymptotes parameters were set to Beta (11, 91), the priors for the intercept or first threshold were set to N (0, 1), and the priors for the second threshold, if applicable, were set to N (–.3, 1). All priors were selected based on preliminary estimation runs. The priors

on the lower asymptote were relatively informative, or restrictive, because the lower asymptote is particularly difficult to estimate, and mis-estimation of this parameter can influence the accuracy of the other parameter estimates (Mislevy, 1986). The priors for the slopes and intercepts were less informative but restrictive enough to minimize uninterpretable results. The lognormal priors forced all specific loadings to be >0. This was a deliberate choice such that the specific factor would be more likely to be something shared by most items on that factor. Otherwise, the factor could be dominated by a single item or pair of items and would have nothing to do with item format.

Neither of the multidimensional models was nested within the other.[1] Non-nested models can be compared using Akaike's Information Criteria (AIC; Akaike, 1987), Bayesian Information Criteria (BIC; Schwarz, 1978), and the sample-size adjusted BIC (SSA-BIC; Sclove, 1987). These indices are based on the –2LL (–2 log-likelihood), with varying penalties for the number of parameters estimated. Lower values indicate better fit, but there is no associated statistical significance test. AIC = –2LL + 2p, where p is the number of parameters estimated. The AIC does not take the sample size (N) into account. Because the AIC tends to favor more complex models with larger sample sizes, the BIC and SSA-BIC increase the penalty for model complexity as N increases. The BIC = –2LL + p (ln(N)). The SSA-BIC replaces N with (N + 2)/24. The AIC, SSA-BIC, and BIC values were 97,192, 97,441, and 97,710, respectively, for the bifactor model and 97,179, 97,364, and 97,563 for the correlated factors model. Thus, the correlated factors model fit slightly better. However, model fit was relatively close and the desired meaning of the scores should also be taken into consideration.

A unidimensional model is nested within both of the multidimensional models. The difference in –2LL for the unidimensional model compared to the bifactor model was $\chi^2(23) = 83.0$, $p < .0001$. The difference in –2LL for the unidimensional model compared to the correlated factors model was $\chi^2(1) = 51.8$, $p < .0001$. This suggests that both multidimensional models fit better than the unidimensional model but this test is problematic because it has a high Type I error rate (Hayashi, Bentler, & Yuan, 2007). Thus, the information criteria are useful for these comparisons as well. The AIC = 97,229, the SSA-BIC = 97,410, and the BIC = 97,607. The AIC suggests that the unidimensional model fit worse than both multidimensional models, but the SSA-BIC and BIC were between the values for the bifactor and correlated factors model. One might reasonably argue from these indices that the subscores would add little to the overall score, as is often

[1]A second-order model would be nested within both the bifactor and correlated factors model, and these models would fit equally well if the data followed a second-order model. Additionally, if the specific factors were allowed to correlate, the correlated factors model would be nested within the bifactor model (Rindskopf & Rose, 1988).

true for achievement tests (Sinharay, 2010; Sinharay et al., 2011), but in many contexts this decision is made by policymakers without regard to model fit.

Score Interpretation. The scale was identified by fixing the latent mean to 0 and the latent standard deviation to 1 for each factor. *Expected-a-posteriori* (EAP) scoring was used, so the standard deviation of the score estimates was less than 1. The lower the reliability, the smaller the standard deviation of the EAP scores because they are pulled in to the mean proportional to the reliability. The squared standard deviation and the squared standard error sum to approximately 1, the latent score variance. Table 1 shows the standard deviation, mean standard error (posterior standard deviation), and reliability of the score estimates. Mean θ scores are not shown because all means were zero when rounded to the nearest .01. Reliability of the EAP θ scores can be estimated as $1 - \frac{s_e^2}{s_\theta^2}$ (Thissen & Orlando, 2001, p. 118; Wainer, Bradlow, & Wang, 2007, p. 76), where s_e^2 is the mean of the squared standard errors and s_θ^2 is the true latent score variance, fixed to 1 in the estimation process. Thus, the reliability is approximately equivalent to the variance of the EAP estimates.

TABLE 1
Summary of EAP Score Estimates ($\hat{\theta}$) for the TIMSS Science Example

Scale	SD of $\hat{\theta}$	Mean SE[1] of $\hat{\theta}$	Reliability
Bifactor model			
General factor	.84	.54	.71
Multiple choice	.36	.93	.13
Constructed response	.50	.86	.25
Correlated factors model[2]			
Multiple choice	.86	.51	.74
Constructed response	.88	.48	.77
Linear composite			
Multiple choice	.85	.52	.72
Constructed response	.87	.49	.76
Predicted observed score[3]			
Multiple choice			.72
Constructed response			.77

[1]More precisely, the value labeled Mean SE was the square root of the mean of the squared standard errors (SE). Because the $\hat{\theta}$s were the EAP estimates, the standard error was the posterior standard deviation (SD).

[2]The estimated correlation between the factors in the correlated factors model was .89. Due to the influence of the correlation in the bivariate prior used to estimate each examinee's vector of $\hat{\theta}$s, the correlation between the EAP score estimates was .975.

[3]Because the metric of the predicted observed scores, τ_s, was different from the other scores' metric, standard errors and standard deviations are not shown.

In this example, score users might be primarily interested in the general factor. The general factor shows the student's overall science proficiency, independent of item format. As shown in Table 1, most of the reliable portion of the variance was due to the general factor. This can also be seen by examining the estimated discrimination parameters in Table 2; a_g was almost always higher than a_s, so the items discriminated better on the general factor. Given that parameters in the logistic metric are 1.7 times parameters in the normal metric, *a*-parameters

TABLE 2
Item Parameter Estimates for the TIMSS Science Example

	Bifactor				Correlated Factors		
Item	a_g	a_s	d	c	a_k	d	c
			Multiple Choice				
1	1.06	.44	−.01	.18	1.09	.00	.18
2	1.59	.45	−.13	.15	1.62	−.14	.15
3	.98	.36	.44	.13	1.01	.45	.12
4	.88	.48	−.34	.13	.94	−.34	.13
5	.27	.33	1.26	.13	.31	1.24	.12
6	1.04	.51	−.77	.11	1.09	−.76	.11
7	.61	.23	−.07	.16	.61	−.01	.13
8	1.34	.45	−.14	.20	1.38	−.14	.20
9	.96	.46	2.22	.12	1.02	2.19	.12
10	1.20	.42	.18	.15	1.23	.20	.14
11	.60	.43	.82	.12	.65	.80	.12
			Constructed Response				
	a_g	a_s	d_1	d_2	a_k	d_1	d_2
12	1.00	.78	2.28	1.40	1.18	2.22	1.36
13	1.07	.45	−.08	−.71	1.15	−.08	−.70
14	.64	.48	.77	−1.03	.75	.76	−1.02
15	1.33	.48	.12	.	1.41	.12	.
16	.54	.36	−1.26	.	.61	−1.25	.
17	.88	.56	1.26	.32	1.00	1.25	.32
18	.80	.73	1.22	.	.97	1.19	.
19	.98	.26	−.03	.	.98	−.03	.
20	.80	.34	.67	.	.84	.67	.
21	.61	.51	−.01	.	.72	.00	.
22	1.22	.42	1.20	.	1.28	1.20	.
23	1.22	.64	1.86	1.07	1.37	1.85	1.07

Items did not appear in this sequence on the test—item numbers are strictly for convenience. a_g is the discrimination on the general factor, a_s is the discrimination on the specific factor (multiple choice or constructed response), a_k is the discrimination on either the multiple choice factor or the constructed response factor for the correlated factors model, d (or d_1 and d_2) is the item easiness, and c is the lower asymptote.

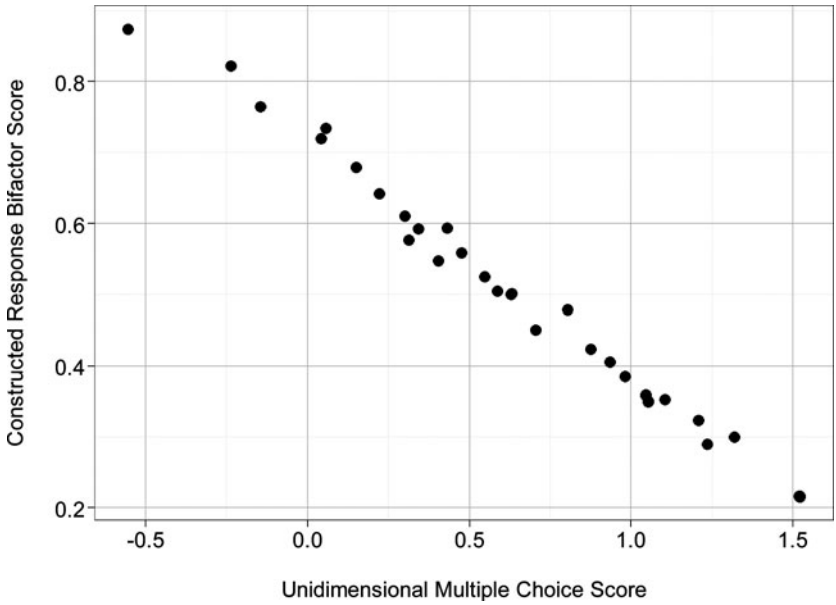

FIGURE 2
Bifactor model estimate of the constructed response trait score (θ) for a subgroup of examinees with identical response patterns on the constructed response items.

around 1.5 or higher indicate reasonable discrimination and a-parameters lower than approximately .8 indicate very little discrimination.

The format-specific scores provided a bit of additional information, although less because the discrimination parameters were lower for the format-specific factors and there were fewer items contributing to each format-specific score. To illustrate the residual nature of the bifactor scores, a group of 36 students was selected with the same response pattern on the constructed response items (101111122122). For this subgroup, the bifactor constructed response θ score was plotted in Figure 2 as a function of the unidimensional θ estimate based only on the multiple choice items. Although all of this subgroup of students did well on the constructed response items, those who also scored well on the multiple choice items received lower constructed response bifactor scores. Van Rijn and Rijmen (2012) and van der Linden (2012) also discussed this phenomenon; answering an item correctly can decrease the examinee's score estimate on one of the factors. In the present context, each observed response on the constructed response scale is modeled as a function of two traits: the general factor and the specific constructed response factor. Holding the response pattern on the constructed response items constant, someone who does not perform very well on the multiple choice items has

a lower level on the general factor. Therefore, the constructed response responses reflect a high level on the constructed response specific factor compensating for the low level on the general factor. However, it should be noted that this relationship is conditional on response pattern; in the sample as a whole the bifactor constructed response θ score was nearly uncorrelated with the unidimensional multiple choice score, but within any subgroup with the same response pattern the correlation was negative.

The bifactor θ score estimates can be contrasted with the correlated factors θ score estimates, also summarized in Table 1. The correlated factors trait scores had lower standard errors and were more reliable because they were more than just the residual part of the responses. For the subgroup plotted in Figure 2, the correlated factors constructed response θ score was plotted in Figure 3, again as a function of the unidimensional θ estimate based only on the multiple choice items. In contrast to Figure 2, those with a high unidimensional score on the multiple choice items received a higher constructed response score. Each θ score in the correlated factors model was essentially a combination of the general and specific factors. Using the correlation between the factors, this model in effect borrows information from the multiple choice factor when estimating the θ score

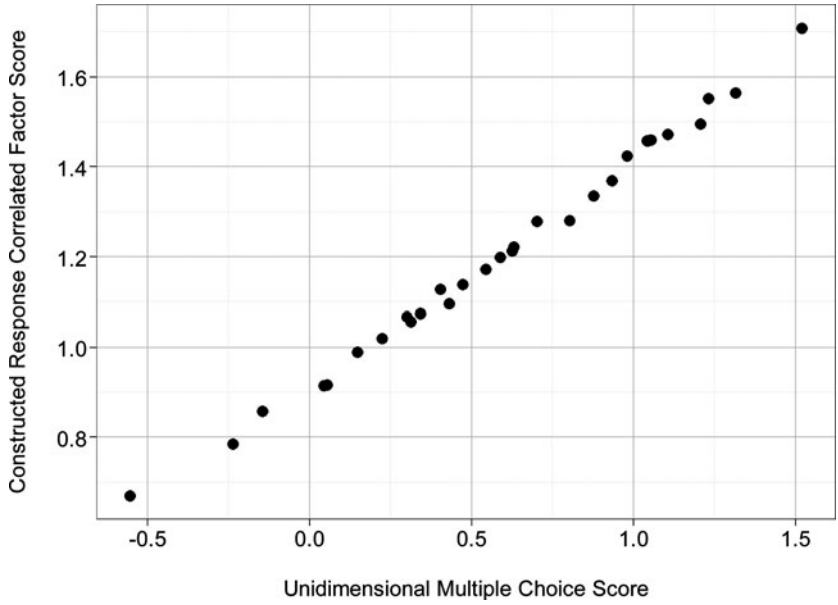

FIGURE 3
Correlated factors model estimate of the constructed response trait score (θ) for a subgroup of examinees with identical response patterns on the constructed response items.

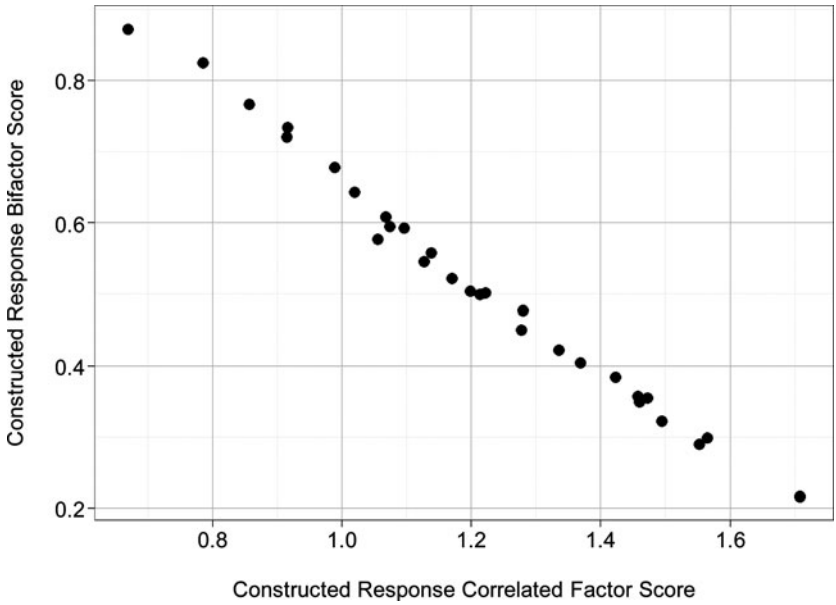

FIGURE 4
Relationship between correlated factors model and bifactor model estimates of the constructed response trait score (θ) for a subgroup of examinees with identical response patterns on the constructed response items.

for the constructed response factor. For a given response pattern on the constructed response items, a student who did well on both scales earned a higher θ score on constructed response than someone who did relatively better on the constructed response items than the multiple choice items. Plotting the bifactor estimates from Figure 2 against the correlated factors estimates from Figure 3, we see a negative relationship in Figure 4 because of these different meanings of the scores. It is critical to note that this negative relationship is conditional on the score pattern for the constructed response items. In the sample as a whole, the bifactor constructed response residuals were positively correlated with the constructed response factor from the correlated factors model, $r = .59$.

Returning to the bifactor scores, two different types of composite score could be calculated to combine information from the general factor and the specific factor if an interpretation akin to the correlated factors model was desired. Wang (1985, as cited in Reckase, 2009, p. 126) used the term *reference composite* to refer to the predominant direction of measurement for the items on a test or subtest. If a subtest mostly measures the general factor, the composite will have a high weight for the general factor. When the general factor is combined with each

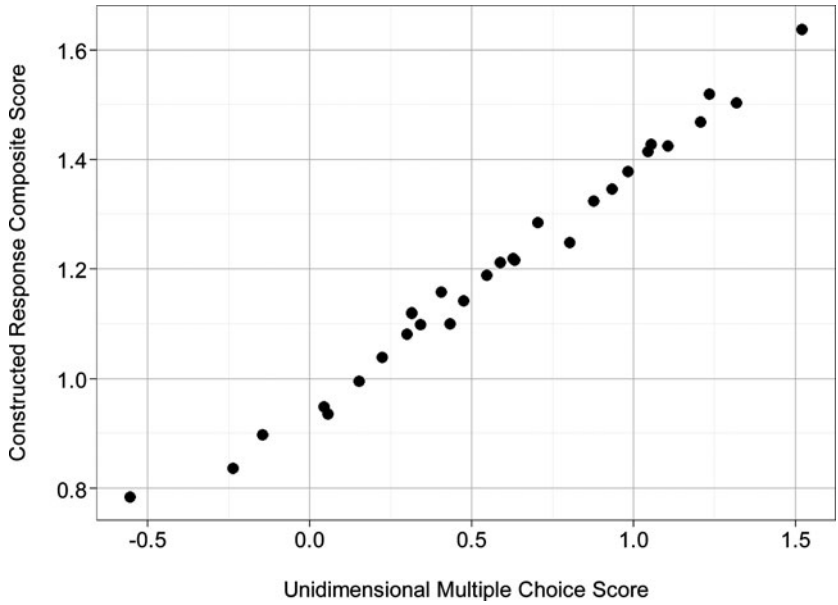

FIGURE 5
Bifactor composite of the general and constructed response trait scores for a subgroup of examinees with identical response patterns on the constructed response items.

specific factor in turn in the direction of the reference composite for the subset of items, the score interpretation is similar to a correlated factors model, as shown in Figure 5. Here, the constructed response composite score = .885 (general factor) + .466 (specific factor). The squared standard error is also a linear combination: $\sigma_E^2 = w_g^2\sigma_{Eg}^2 + w_s^2\sigma_{Es}^2 + 2w_g w_s \sigma_{Esg}$, where σ_E^2 is the error variance, w_g and w_s are the weights for the general and specific factors respectively (.885 and .466), σ_{Eg}^2 and σ_{Es}^2 are the error variances for the general and specific factors, and σ_{Esg} is the error covariance. The mean standard error for the constructed response composite score was .49, for a composite reliability of .76. The mean reliability for the multiple choice composite score was .72.

As described in the introduction, an alternative composite score, τ_s, the predicted observed score, could be estimated. A point estimate of τ_s for subscale s could be defined as $\tau_s = \sum_{i \in s} P_i(\hat{\theta}_g, \hat{\theta}_s)$ for dichotomous items. But to estimate the standard error and score reliability, one must not use the point estimates of $\hat{\theta}_g$ and $\hat{\theta}_s$ but the posterior distributions (see Appendix). After approximating the posterior distribution as a function of both θ_g and θ_s, marginalized over the other

specific dimensions, τ_s can be evaluated at each pair of quadrature points.[2] The mean of the distribution[3] is the estimate of τ_s and the standard deviation is the estimated standard error of τ_s. Using this procedure, the estimated reliability for the multiple choice composite was .72 and the estimated reliability for the constructed response composite was .77. Although these composites were on a different metric, they were correlated .997 and .991 with the linear composites. The Spearman correlations, which consider only rank order and not linearity, were .99995 and .99698, respectively. Additionally, when limited to the subset of examinees depicted in the figures, the nonlinearity was not visually discernible. Therefore, no additional figures are shown for this second composite score.

Using either method of composite scores, the score estimates and reliability values are similar to those from the correlated factors model, but this procedure makes it clear that most of the reliable portion of the combination was due to the general factor. The composite scores allow for the use of the bifactor model with an interpretation similar to the correlated factors model if desired.

In summary, the responses to the off-scale items have an impact on the $\boldsymbol{\theta}$ trait score estimates. In the bifactor model, an examinee's correct responses can be due either to a high level on the general factor or a high level on the specific factor. If the examinee scored well on the off-scale items, then the score on the on-scale items can be attributed to a high level of the general factor, thus pulling the estimate of the trait score on the specific factor down. In the correlated factors model or the composite of the general and specific bifactor scores, the effect of the off-scale items is to pull the trait score for factor *k* closer to the trait scores on the other factors so a high score on the off-scale items can increase the estimate of the trait score.

Civic Responsibility Example

The Civic Responsibility Behavior Questionnaire (Markle, 2009) was designed to assess the development of civic responsibility in university students. The dataset included 1128 incoming students as well as 1258 students halfway through their second year; six students were eliminated from the analysis because they answered fewer than half the items, leaving a total of 2380 students. These students were randomly selected from within the cohorts of students required to participate in assessment. The instrument contained 22 items on four scales: Civic Behavior (5 items), Political Behavior (6 items), Social Behavior (6 items), and Values (5 items). All items had a 5-option response scale, ranging from *Never or Rarely* to

[2]For this demonstration, there were 49 points for each dimension. Thus, the distribution of τ was approximated on 2401 points.

[3]Due to nonlinearity, the mean of the distribution will not be precisely the value evaluated at $\hat{\theta}_g, \hat{\theta}_s$.

Frequently for the Behavior scales and *Strongly Disagree* to *Strongly Agree* for the Values scale.

Item Parameter Estimation. The bifactor and correlated factor models were again run in flexMIRT (Cai, 2012). For the purpose of this tutorial, no efforts were made to assess whether the subscales actually represented the underlying factor structure. That is an important component of instrument development,[4] but the focus here is not on instrument development but on interpreting the resulting scores after the model has been confirmed. Typically, one would posit multiple models based on the theory behind the instrument, compare the fit of the models, perhaps additionally make some small modifications to either the model or the instrument based on the empirical results, and cross-validate with another sample if changes were made (Bandalos & Finney, 2010; Boomsma, 2000; Mueller, 1997). For this exposition on score interpretation, however, it is assumed these steps were completed previously. Thus, only a single factor configuration was modeled.

The priors for the slopes were N(1.7, 1) for the general factor and for each factor in the correlated factors model, and logNormal (–.2, .5) for the specific factors. These priors were moderately informative and strong enough to prevent implausible values. The multidimensional extensions of the graded response model were used for both the bifactor and correlated factor models. The number of quadrature points was reduced to 22^4 for the correlated factors model because the estimation would not proceed with greater numbers.

The AIC, SSA-BIC, and BIC values were 130,533, 130,876, and 131,369, respectively, for the bifactor model and 131,976, 132,278, and 132,711 for the correlated factors model. Thus, the bifactor model fit slightly better, but the correlated factors scores again provide a useful contrast to help understand what the bifactor scores mean. Both models fit significantly better than a unidimensional model ($\chi^2\,(22) = 8958.6$ for the comparison with the bifactor model and $\chi^2\,(6) = 7483.3$ for the comparison with the correlated factors model, $p < .0001$ for both comparisons). AIC = 139,448, SSA-BIC = 139,733, BIC = 140,083, all suggesting worse fit for the unidimensional model.

Score Interpretation. Table 3 provides the standard deviations of the trait score estimates, the average SE, and the marginal reliability. Again, means for all score estimates were zero when rounded to the nearest .01. For the bifactor model, the Civic Behavior trait estimates had low reliability. These items had among the highest discriminations on the general factor, but low to moderate discriminations on the specific trait (Table 4); these items measured the general factor well but did

[4]For an overview of the importance of model testing and comparisons and how to conduct them, see Kline, 2010, Chapters 4, 8, and 9.

TABLE 3
Summary of Score Estimates for the Civic Responsibility Example

Scale	SD of $\hat{\theta}$	Mean SE[1] of $\hat{\theta}$	Reliability
Bifactor model			
General factor	.93	.31	.90
Civic Behavior	.60	.79	.36
Political Behavior	.82	.58	.67
Social Behavior	.95	.34	.89
Values	.88	.48	.77
Correlated factors model[2]			
Civic Behavior	.96	.30	.91
Political Behavior	.95	.31	.90
Social Behavior	.95	.32	.90
Values	.92	.40	.84
Linear combination			
Civic Behavior	.94	.29	.91
Political Behavior	.94	.30	.91
Social Behavior	.95	.32	.90
Values	.91	.40	.84
Predicted observed scores[3]			
Civic Behavior			.91
Political Behavior			.91
Social Behavior			.89
Values			.85
Estimated correlations[2]	Civic	Political	Social
Civic Behavior			
Political Behavior	.93		
Social Behavior	.23	.17	
Values	.48	.44	.36

[1] More precisely, the value labeled Mean SE was the square root of the mean of the squared standard errors (SE).

[2] The values in the table are the estimated latent correlations. Due to the influence of the correlation in the multivariate prior, the observed correlation between the expected-a-posteriori score estimates ranged from .19 between Political and Social Behavior to .97 between Civic and Political Behavior.

[3] Because the metric of the predicted observed scores, τ_s, was different from the other scores' metric, standard errors and standard deviations (SD) are not shown.

not measure much beyond that. As would be expected from the high general factor loadings, this scale had much higher reliability in the correlated factors model. Social Behavior and Values, in contrast, had higher bifactor reliability than the other subscales because they had higher discriminations on the specific traits. The reliability of the Social scale was comparable on both the bifactor and correlated factors model because these items had low discriminations on the general factor

TABLE 4
Item Parameter Estimates for the Civic Responsibility Example

	Bifactor						Correlated Factors				
Item	a_g	a_s	d_1	d_2	d_3	d_4	a_k	d_1	d_2	d_3	d_4
					Civic Behavior						
36	3.41	1.56	5.34	2.23	−.69	−3.62	3.15	4.77	2.04	−.52	−3.10
38	3.93	1.63	6.18	2.87	−.56	−3.82	3.62	5.54	2.64	−.40	−3.28
40	1.97	.17	3.46	1.57	−.29	−2.11	1.90	3.46	1.61	−.22	−2.03
42	2.21	.08	2.23	.08	−2.07	−3.89	1.89	2.12	.14	−1.85	−3.54
48	1.09	.10	2.06	.43	−1.03	−2.60	1.05	2.08	.46	−.99	−2.55
					Political Behavior						
37	2.34	.09	2.37	.45	−1.47	−3.50	2.01	2.25	.48	−1.30	−3.18
41	2.26	.23	3.09	1.00	−1.07	−2.80	2.19	3.13	1.06	−.99	−2.72
43	1.05	1.13	1.10	.47	−.41	−1.47	1.11	.97	.40	−.36	−1.28
44	1.93	1.01	1.87	.42	−1.66	−3.55	2.04	1.86	.45	−1.56	−3.40
45	3.18	3.42	5.16	2.35	−.60	−3.75	2.05	2.81	1.30	−.28	−1.99
46	2.31	.98	3.33	1.47	−.58	−2.61	2.46	3.39	1.54	−.50	−2.54
					Social Behavior						
33	1.06	.00	3.31	2.42	1.13	−.45	.24	3.12	2.29	1.11	−.30
34	.67	1.75	2.84	1.15	−1.24	−3.16	1.87	2.86	1.17	−1.22	−3.14
35	.74	2.85	3.54	1.12	−1.82	−4.40	2.96	3.57	1.15	−1.81	−4.40
39	.85	3.18	4.47	1.96	−1.12	−3.60	3.31	4.51	2.00	−1.09	−3.59
47	.82	4.54	5.94	2.76	−1.48	−5.00	4.31	5.62	2.62	−1.37	−4.70
49	.72	.46	2.49	1.02	−.51	−1.83	.59	2.39	.98	−.47	−1.70
					Values						
51	1.00	1.67	5.53	3.52	.81	−1.80	1.95	5.58	3.56	.84	−1.78
52	1.04	2.49	7.17	4.76	1.36	−2.10	2.46	6.79	4.50	1.30	−1.94
53	.83	1.61	5.12	3.45	1.25	−1.60	1.79	5.12	3.46	1.27	−1.57
54	1.49	1.89	6.08	3.67	.37	−2.75	2.47	6.25	3.79	.42	−2.78
55	.86	1.44	6.11	3.97	1.67	−1.13	1.67	6.14	3.99	1.69	−1.10

a_g is the discrimination on the general factor, a_s is the discrimination on the specific factor, the a_k are the discrimination parameters on the factors of the correlated factors model, and $d_1 - d_4$ are the item thresholds.

in the bifactor model and correspondingly low correlations with the other factors in the correlated factors model.

The Values scale was chosen to illustrate trait score interpretations. One of the more common response patterns, selected by 52 respondents, was 34434. If the Values scale were scored separately, each of these respondents would receive the same θ score, somewhat below average compared to the total group. Trait score estimates on the Values scale are shown in Figure 6. The bifactor and correlated factor estimates are plotted directly against each other. In this subgroup with the same response pattern on the Values items, those who scored higher on

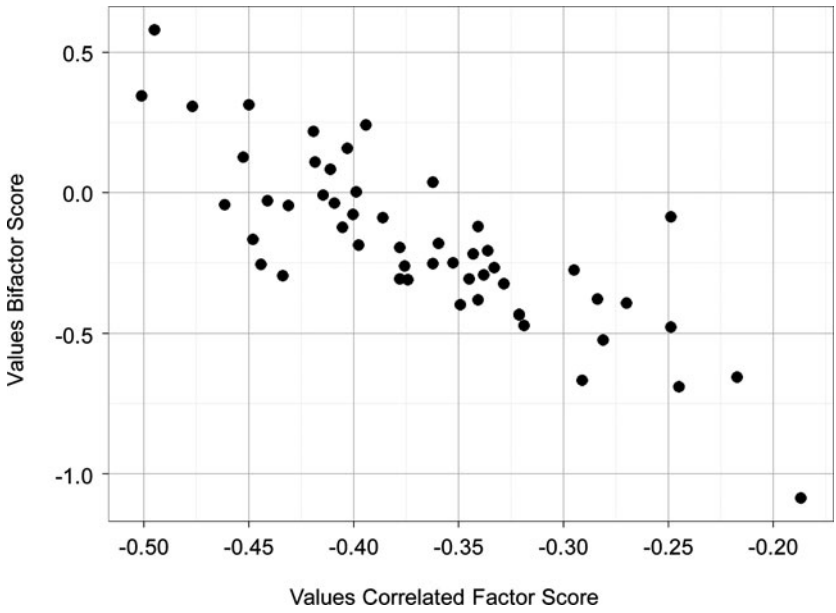

FIGURE 6
Relationship between bifactor model and correlated factors model estimates of the Values trait score (θ) for a subgroup of examinees with identical response patterns on the Values items.

the general factor had lower Values bifactor θ scores but higher Values correlated factor model θ score estimates. Thus, the bifactor and correlated factor trait estimates were negatively correlated. It should be emphasized that this relationship only holds when the subscale response pattern is held constant. In the sample as a whole, there was a positive relationship between θ estimates from the two models, $r = .86$.

A linear composite of the general and specific scores was also calculated, with weights of .49 for the general factor and .87 for the Values factor. τ_s was also calculated for the items on the Values scale. The two sets of composites were correlated .93, which was not as high as the correlation between the bifactor composites in the TIMSS example. The differences were not due just to nonlinearity; the Spearman correlation was only .89, indicating the ordering of examinees was not perfectly consistent. The correlated factor model scores were correlated .997 with the linear composite and .93 with τ_k. The reliabilities of both composites, .84 and .85, respectively, were also very similar to the correlated factors model reliability. The same similarities between the composites and the correlated factors model were observed for the other subscales, as shown in Table 3. Again,

the composite scores allow for an interpretation similar to the correlated factors model, with the computational simplicity of the bifactor model and possibly better model fit.

The weights of the general and specific factors varied by subscale for the linear composite, and by item for τ_k. Looking at Table 4, the items on the Social Behavior subscale had relatively low *a*-parameters on the general factor and high *a*-parameters on the specific factor. For the linear composite, the weights were .24 for the general factor and .97 for the specific factor. Civic Behavior, in contrast, would have weights of .95 for the general factor and .32 for the specific factor because these items had high *a*-parameters on the general factor and lower *a*-parameters on the specific factor. These composites make explicit a process also underlying the correlated factors model: The more correlated the factors, the less unique information is contributed by the subscale.

In this civic responsibility scale, the bifactor specific factor trait scores would be useful for focusing on the unique contribution of each subscale, beyond civic responsibility in general. For example, holding constant self-reported Behaviors constant, which respondents profess higher Values? In contrast, the correlated factors model or a composite of general and specific would be useful when examining overall standing on each scale.

DISCUSSION AND CONCLUSIONS

As these examples have illustrated, the interpretation of the specific factor score estimates from the bifactor model requires an understanding of their residual properties. Ultimately, when the bifactor model is applied, the way in which the scores will be used should be a major consideration in deciding whether the specific factor scores or weighted combinations of the general and specific factors should be reported.

The general and specific scores approach is obviously useful if the scores are used in a latent regression model. In a latent regression model, the individual student scores would not need to be estimated, but the interpretation is the same as it would be for individual student score reports. As discussed by Chen, West, and Sousa (2006), using the bifactor model shows the unique contribution of each of the specific factors "over and above" the general factor. They noted that factors from the correlated factors model may contribute little individually to prediction of an external criterion due to multicolinearity; this criticism would apply to the linear combination of general and specific factors as well.

Decisions are more difficult in contexts where scores are to be reported to students or clients, or to their teachers or counselors. If discrimination parameters for both the general factor and the specific factor are high, one could reasonably

argue for either reporting the general and specific factor score estimates, or reporting a weighted combination, depending on the types of interpretations one wanted to make. In this context, it is important to distinguish between scores that reflect the examinee's strengths and weaknesses after controlling for the general factor, compared to scores that reflect the examinee's overall standing on each scale. The specific factor scores are useful for the first purpose, but the weighted combinations are more useful for the second purpose. The weighted combination score probably better matches most score users' innate understanding of what a subscore represents. However, if separate general and specific factor score estimates appear more appropriate for score users' needs, score reports and auxiliary materials should be written to guide proper interpretation of the residual nature of the factor scores.

If loadings are high on the general factor but low on the specific factor, it might be most defensible to report the general factor scores alone. Reise, Moore, and Haviland (2010) and Reise, Morizot, and Hays (2007) contended that observed (number correct or scaled transformations of number correct) subscale scores are misleading when loadings on the general factor are high and loadings on the specific factors are low. They reasoned that most of the reliable portion of the subscale was due to the general factor and thus the subscale scores were largely redundant. This argument could be extended to scores calculated as a weighted combination of the general factor and the specific factor, although this approach at least makes it clear how little unique variance is contributed by the specific factor.

If policy considerations make some type of subscale score necessary even when loadings on the specific factor are low, there are psychometric issues with either the specific factor scores or the weighted combination of general and specific factors. The specific factors are likely to be unreliable, and reporting the low reliability along with the scores will not keep users from ignoring the low reliability and interpreting the scores regardless. The weighted combination of the general and specific factor will be more reliable and have a smaller standard error. However, users may overinterpret any small differences between subscores, which are entirely due to the less reliable specific factors. Each specific factor score estimate provides little information beyond that provided by the general factor estimate. Sinharay and colleagues (2011) advised: "Whenever subscores are provided, provide evidence of adequate reliability, validity, and distinctness of the subscores" (p. 36). Trait scores estimated as a function of a strong general factor and a weak specific factor may be quite reliable, but they are not very distinct and will likely not demonstrate differential validity. The same caveat extends to factor scores from a correlated factors model.

In summary, when loadings on the specific factors are high, factor score estimates for the specific scores can be meaningful as long as score users are carefully educated about the fact that each subscale reflects information above and beyond the skill or trait reflected in the general score. When loadings on the

specific factors are low, only the general factor score carries a reliable interpretation. If subscores must be reported, a weighted combination of the general and specific factor may be less prone to misuse than the unreliable specific factor scores alone.

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317–332.

Armon, G., & Shirom, A. (2011). The across-time associations of the five-factor model of personality with vigor and its facets using the bifactor model. *Journal of Personality Assessment*, *93*, 618–627.

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93–114). Florence, KY: Routledge Education.

Bear, G. G., Gaskins, C., Blank, J., & Chen, F. F. (2011). Delaware School Climate Survey—Student: Its factor structure, concurrent validity, and reliability. *Journal of School Psychology*, *49*, 157–174.

Betts, J., Pickart, M., & Heistad, D. (2011). Investigating early literacy and numeracy: Exploring the utility of the bifactor model. *School Psychology Quarterly*, *26*, 97–107.

Bock, R. D., & Gibbons, R. (2010). Chapter 7: Factor analysis of categorical item responses. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 155–184). New York, NY: Routledge.

Bock, R. D., Gibbons, R., & Muraki, E. (1998). Full-information factor analysis. *Applied Psychological Measurement*, *12*, 261–280.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling*, *7*, 461–483.

Brown, A. R., Finney, S. J., & France, M. K. (2011). Using the bifactor model to assess the dimensionality of the Hong Psychological Reactance Scale. *Educational and Psychological Measurement*, *71*, 170–185.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

Cai, L. (2012). flexMIRT^{TM} version 1.88: A numerical engine for multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized item bifactor analysis. *Psychological Methods*, *16*, 221–248.

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

Gibbons, R. D., Bock, D. R., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., … Stover, A. (2007). Full-information item bi-factor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4–19.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika*, *57*, 423–436.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Hayashi, K., Bentler, P. M., & Yuan, K.-H. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling*, *14*, 505–526.

Kim, S.-H., Sherry, A. R., Lee, Y. S., & Kim, C.-D. (2011). Psychometric properties of a translated Korean adult attachment measure. *Measurement and Evaluation in Counseling and Development*, *44*, 135–150.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Markle, R. E. (2009, October). *How do you act like a citizen: A confirmatory factor analysis of the Civic Responsibility Behavior Questionnaire*. Paper presented at the Northeastern Educational Research Association Conference, Rocky Hill, CT.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177–195.

Mueller, R. O. (1997). Structural equation modeling: Back to basics. *Structural Equation Modeling*, *4*, 353–369.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*, 73–90.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401–412.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York, NY: Springer.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer-Verlag.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*, 361–373.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544–559.

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.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*, 361–372.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, *23*, 51–67.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333–343.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150–174.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29–40.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen and H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah, NJ: Lawrence Erlbaum Associates.

van der Linden, W. J. (2012). On compensation in multidimensional response modeling. *Psychometrika*, *77*, 21–30.

van Rijn, P. W., & Rijmen, F. (2012). *A note on explaining away and paradoxical results in multidimensional item response theory* (Report No. ETS RR-12-13). Princeton, NJ: ETS.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement*, *47*, 339–360.

APPENDIX: FULL INFORMATION IRT

The bifactor model is especially practical when full information estimation methods are applied. Full information estimation considers the likelihood of the entire response string, as opposed to limited information methods which utilize only pairwise information, such as correlations or bivariate proportion correct. Full information is widely used in item response theory (IRT), where MML (maximum marginal likelihood or marginal maximum likelihood) is a common estimation technique. In MML, the likelihood is evaluated at a series of points, called quadrature points, over the distribution of ability. If one uses 10 quadrature points for each factor, one needs 10^K quadrature points, where K is the number of factors. However, for the bifactor model, software programmers can take advantage of the fact that each item loads only on the general factor and up to one specific factor, so no more than q^2 quadrature points are needed, where q is the number of points per factor. This makes the bifactor model with K factors, one general factor and *K-1* specific factors, more computationally feasible than the correlated factors model with *K-1* factors (Bock & Gibbons, 2010; Cai, 2010; Cai et al., 2011; Gibbons & Hedeker, 1992; Rijmen, 2010).

In MML estimation, the likelihood of each response pattern is marginalized (averaged) over the trait distributions. In the more general multidimensional model with K factors, the marginal likelihood of response pattern $\mathbf{x}_j$ is (Gibbons & Hedeker, 1992):

$$P(\mathbf{x} = \mathbf{x}_j) = \int \cdots \int L_j(\boldsymbol{\theta})\, g(\theta_1) \cdots g(\theta_K)\, d\theta_1 \cdots d\theta_K, \tag{5}$$

where $L_j(\boldsymbol{\theta})$ is the likelihood of the response pattern j given the current estimates of the item parameters, and $g(\boldsymbol{\theta})$, the density of $\boldsymbol{\theta}$, is typically assumed multivariate normal although it could be estimated empirically. The likelihood of response pattern j given $\boldsymbol{\theta}$ is the product of the likelihood of each item response given $\boldsymbol{\theta}$: $L_j = \prod_i P_i(\boldsymbol{\theta})$, where $P_i(\boldsymbol{\theta})$ is calculated from Equation 1 or 4. The integration in Equation 5 is approximated by evaluating the likelihood at a series of quadrature points. This is why q^K quadrature points are generally needed for multidimensional MML estimation. For more details on MML estimation for the general case, see Bock, Gibbons, and Muraki (1998) or Muraki and Carlson (1995).

For the bifactor model, however, only the general factor and one of the *K–1* specific factors are involved in the likelihood for any one item. For any specific factor s, one first calculates L_{js} only from the items which load on factor s, as a function of θ_1 and θ_s instead of the entire $\boldsymbol{\theta}$: $L_{js} = \prod_{i \in s} P_i(\theta_g, \theta_s)$. The function is integrated over $g(\theta_s)$ before the marginal L_{js} are multiplied to obtain L_j as a function of θ_g. L_j can then be marginalized over θ_g:

$$P(\mathbf{x} = \mathbf{x}_j) = \int \prod_{s=1}^{S} \left[\int L_{js}(\theta_g, \theta_s) g(\theta_s) d(\theta_s) \right] g(\theta_g) d(\theta_g). \tag{6}$$

Thus, only q^2 quadrature points are needed for MML estimation of the bifactor model.

At each step of the estimation, the item parameters that maximize the marginal likelihood in Equation 4 are estimated. For details about the EM algorithm used in this maximization, see Gibbons and Hedeker (1992).

Trait Score Estimates

After the item parameters have been estimated, parts of Equation 6 can be used to estimate the bifactor trait scores $\boldsymbol{\theta}$. For the general trait, θ_{g}, the posterior probability of response pattern j is marginalized over all of the specific traits, weighted by the density of θ_{g}, and divided by the marginal likelihood of the response pattern, to give the posterior probability of θ_{g}. The mean of this posterior is the *expected a-posterior* (EAP) estimate: $\hat{\theta}_g = \int \theta_g \mathrm{P}(\theta_g|\mathbf{x} = \mathbf{x}_{\mathrm{j}})d(\theta_g)$. The standard deviation of the posterior distribution is an estimate of the standard error. For specific factor s, the integration is instead over each of the other specific factors, followed by the general factor, to find the posterior distribution of θ_{s}.