Evaluation of the Linear Composite Conjecture for Unidimensional IRT Scale for Multidimensional Responses

Tyler Strachan\textsuperscript{1,*}, Uk Hyun Cho\textsuperscript{1}, Terry Ackerman\textsuperscript{2}, Shyh-Huei Chen\textsuperscript{4}, Jimmy de la Torre\textsuperscript{3}, and Edward H. Ip\textsuperscript{4}

Abstract

The linear composite direction represents, theoretically, where the unidimensional scale would lie within a multidimensional latent space. Using compensatory multidimensional IRT, the linear composite can be derived from the structure of the items and the latent distribution. The purpose of this study was to evaluate the validity of the linear composite conjecture and examine how well a fitted unidimensional IRT model approximates the linear composite direction in a multidimensional latent space. Simulation experiment results overall show that the fitted unidimensional IRT model sufficiently approximates linear composite direction when correlation between bivariate latent variables is positive. When the correlation between bivariate latent variables is negative, instability occurs when the fitted unidimensional IRT model is used to approximate linear composite direction. A real data experiment was also conducted using 20 items from a multiple-choice mathematics test from American College Testing.

Keywords

linear composite, unidimensional scale, item response theory, multidimensional item response theory

It is generally recognized that a test, despite the original intention to assess a single targeted construct, often inadvertently includes secondary constructs that are peripheral to the assessment but could influence the interaction between the test taker and the test items. Humphreys (1982) for example, coined the term “systematic heterogeneity” to describe the characteristic of a test that is sufficiently broad to include minor contents in order to become a meaningful and valid test. As argued by many others (Hambleton & Swaminathan, 1985; Reckase, 1979,

\textsuperscript{1}Educational Research Methodology, University of North Carolina at Greensboro, Greensboro, NC, USA
\textsuperscript{2}University of Iowa, Iowa City, IA, USA
\textsuperscript{3}Department of Education, The University of Hong Kong, Pokfulam, Hong Kong
\textsuperscript{4}Department on Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA

Corresponding Author:
Tyler Strachan, Educational Research Methodology, University of North Carolina at Greensboro, 1400 Spring Garden Street, Greensboro, NC 27412, USA.
Email: tstrachan007@rocketmail.com
In educational and psychological testing, a test that is designed to measure a single construct—psychological, educational, or otherwise, is likely to involve other minor constructs in practice. Thus, there exists a dilemma between the psychometric desire for assessing a single construct versus the need for a test to function meaningfully as a valid instrument. Ip (2010) called this the dimensionality versus validity dilemma.

Yet another scenario the dimensionality versus validity dilemma could arise is when a test is designed to measure multiple constructs. However, because of practical constraints, it is desirable to report a unidimensional score. Researchers have to determine if the test is sufficiently unidimensional for reporting purpose.

One of the most commonly used psychometric models for measuring a single construct is the unidimensional item response theory (UIRT). Because of the recognition about the potential presence of multiple dimensions in a test, multidimensionality IRT (MIRT) has also been increasingly used to analyze educational and psychological data. The trend of MIRT is also partly driven by the emergence of stable MIRT software programs (Han & Paek, 2014). With the increasing availability of mature software tools in MIRT to analyze data, the balance in the discussion about the dimensionality versus validity dilemma is shifting. As simulation and analytic tools become more efficient and less costly, this dilemma now can be thoroughly investigated from both MIRT and UIRT perspective.

In this paper we take advantage of available MIRT tools to examine a long-standing conjecture about the characteristics of a fitted UIRT model to multidimensional responses—that is, the fitted UIRT to multidimensional data represents a linear composite of the dimensions present in a test (Reckase, 1979; Reckase et al., 1986; Wang, 1987; Luecht & Miller, 1992; Camilli, 1992; Zhang & Stout, 1999a,b; Kahraman & Thompson, 2011; Ackerman et al., in press). Specifically, using the probit (normal ogive) form, the linear composite conjecture (LCC, Wang, 1987, p.30) states that when item responses are generated from an MIRT model (McDonald, 1997)

\[ P(Y_i | \theta_i) = \Phi(A\theta_i + d) \]

then the fitted UIRT to the data represents a linear composite that points in the direction of the first eigen vector of the matrix \(A^T A\) (Wang, 1987, p.30). Here, \(Y_i\) represents the vector of (dichotomous) response from the \(i^{th}\) person \((i = 1,..,I)\) to the \(j^{th}\) item \((j = 1,..,J)\), \(\theta_i\) the vector of abilities in the \(m\)-dimensional latent space \((m > 1)\), \(A\) the \(J \times m\) matrix of discrimination parameters in a compensatory MIRT, \(d\) the vector of intercept item parameters, and \(\Phi(.)\) the cumulative normal distribution function. This conjecture is important because if it is valid it will allow a simplified mechanism for understanding the fitted UIRT to multidimensional data. A researcher will be able to anticipate the behavior of the UIRT with varying underlying MIRT parameters. As a result, sensitivity analysis can be efficiently conducted for assessing how specific changes to candidate test items could affect the composite direction. The information derived from the LCC can also be used to exclude items that do not fit into what Ackerman (1992) called the validity sector, which refers to the limit of the extent to which the secondary dimensions are allowed into the test. Despite its importance, the LCC has not been comprehensively evaluated, although a substantial literature about the robustness of UIRT to model misspecification does exist (Drasgow & Parsons, 1983; Harrison, 1986; Ackerman, 1994, Ackerman, 1989; Junker & Stout, 1994; Zhang & Stout, 1999; Reise et al., 2014). In an unpublished doctoral dissertation, Kim (1994) compared Wang’s LCC with some other conjectures of composites—all linear in nature, using simulations. However, the conditions for the simulation were limited. It is interesting to note that while authors expressed doubts about linearity of the LCC (Junker, 1994), preliminary simulation work from Junker’s group (Kim, 1994, p. 97)
seemed to show consistency between the fitted UIRT and Wang’s linear composite direction. We will further discuss the linearity issue in the Discussion section.

In the current paper we evaluated the LCC using both simulated and real data. In our simulation experiment, the evaluation of the conjecture follows these steps: 1. Generate response data from an MIRT; 2. Compute the composite direction based on the true MIRT item parameters per Wang’s LCC; 3. Mathematically derive the conjectured fitted UIRT item parameters and person ability using multivariate normal theory (Zhang & Wang, 1998); 4. Fit UIRT to the data; and 5. Compare both item parameter and ability estimates of the fitted UIRT to the conjectured model in Step 3. In the real data analysis, we used a confirmatory MIRT model as the reference model. Estimated MIRT parameters were used to obtain the LCC model, and model parameters between the LCC model and the fitted UIRT were directly compared.

The remainder of the paper is organized as follows. First, we give background about the computation of the linear composite, then we describe the simulation experiment and the real data analysis. We focus on the two-dimensional MIRT. Finally, we provide a discussion.

**Computing a Linear Composite**

Assume that the item response function can be denoted by the two-dimensional compensatory logistic model (Reckase, 2009) and ignoring the person index

\[ P_j(Y = 1 | \theta_1, \theta_2) = \frac{1}{1 + \exp[-a_{j1}\theta_1 - a_{j2}\theta_2 - d_j]} \]  

where \( a_{j1} \) and \( a_{j2} \) are slope parameters for the \( j \)th item and \( d_j \) is a scalar parameter related to intercept of the \( j \)th item. It is also assumed that the latent variables \( \theta_1 \) and \( \theta_2 \) follows a multivariate normal distribution. Without loss of generality, it’s assumed that \( \theta_1 \) and \( \theta_2 \) are standardized, that is, \( \theta_1, \theta_2 \sim N(0, \Sigma) \), where \( \Sigma = (\sigma_{nn}) \) is a \( 2 \times 2 \) positive definite matrix with unit diagonal elements. Under these two assumptions, the true score of a single reported score may be related to the latent variables through a linear composite (Zhang & Wang, 1998). A linear composite \( \theta_\alpha \) of the latent variables is defined to be a standardized linear combination of \( \theta_1 \) and \( \theta_2 \) such that

\[ \theta_\alpha = \alpha^T \theta = \alpha_1 \theta_1 + \alpha_2 \theta_2 \]  

where \( \alpha = (\alpha_1, \alpha_2) \) is a vector of weights with non-negative \( \alpha \)'s and the following constraint must be satisfied

\[ \alpha^T \Sigma \alpha = 1 \]  

where \( \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \) and the vector \( \alpha \) represents the direction of composite \( \theta_\alpha \) in the two-dimensional latent space. To obtain the elements of vector \( \alpha \), let’s first denote \( A \) as the matrix of two-dimensional item discrimination parameters. The eigen decomposition of \( A^T A \) is computed to obtain the eigenvector of the first eigenvalue \( v_{\lambda_1} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \). Once \( v_{\lambda_1} \) is obtained the first weight \( \alpha_1 \) is computed as follows

\[ \alpha_1 = \frac{v_1}{\|v_{\lambda_1}\|} \]
With algebraic manipulation, \( \alpha_2 \) can be solved via equation (4) which results in the following quadratic equation

\[
\alpha_2^2 + 2\rho \alpha_1 \alpha_2 + \alpha_1^2 - 1 = 0
\]

where the coefficients of the quadratic equation are \( a = 1 \), \( b = 2\rho \alpha \), and \( c = \alpha_1^2 - 1 \). If \( \rho = 0 \), the weight \( \alpha_2 \) can simply be obtained by:

\[
\alpha_2 = \sqrt{1 - \alpha_1^2}
\]

The marginal item response function with respect to a linear composite is represented as a unidimensional response function under the assumptions of multivariate normal \( \theta_1 \) and \( \theta_2 \) and two-dimensional compensatory logistic model

\[
P_j(Y = 1|\theta_a) = \frac{1}{1 + \exp[-a_j^* \theta_a - d_j^*]}
\]

where

\[
a_j^* = \left(1 + \sigma_j^{*2}\right)^{\frac{1}{2}} a_j^T \sum \alpha
\]

\[
d_j^* = \left(1 + \sigma_j^{*2}\right)^{\frac{1}{2}} d_j
\]

and

\[
\sigma_j^{*2} = a_j^T \sum a_j - \left(a_j^T \sum \alpha\right)^2
\]

Recall that \( \alpha^T \Sigma \alpha = 1 \) is assumed. Figure S1 in supplementary materials illustrates the linear composite direction in a two-dimensional latent space.

**Simulation Experiment**

The purpose of the simulation experiment was to examine how similar the linear composite scale, \( \theta_a \), obtained from the generated two-dimensional MIRT (2D-MIRT) was to the estimated unidimensional IRT (UIRT) scale, \( \hat{\theta} \). A fully crossed factorial simulation design was run in R software (R Core Team, 2017). The mirt package (Chalmers, 2012) was the software of choice for generating response data, item calibrations, and scoring respondents. Several factors were manipulated in the simulation experiment including sample size, test length, correlation for \( \theta_1 \) and \( \theta_2 \), and correlation for \( \alpha_1 \) and \( \alpha_2 \). A sample size of \( N = 250, 500, \) and 1,000 were randomly generated from a multivariate normal distribution with \( \mu = (0, 0) \) and covariance matrix

\[
\Sigma = \begin{pmatrix}
1 & \rho \\
\rho & 1
\end{pmatrix}
\]

which consisted of unit diagonal elements and correlation, \( \rho \), between the latent variables \( \theta_1 \) and \( \theta_2 \). The \( \rho \) between the latent variables was manipulated at different levels including \(-0.8, -0.4, 0.0, 0.4, 0.8, \) and 1 (for UIRT model). A test length of \( J = 40, 80, 100, \) and
150 items were randomly generated where \( a_1 \) and \( a_2 \) were either positive or negatively correlated. The \( a_1 \) and \( a_2 \) were generated from separate uniform distributions such that when correlating the item vectors of \( a_1 \) and \( a_2 \) (across \( J \) items) the empirical correlation (\( r \)) was either negative (e.g., the \( a_1 \) and \( a_2 \) item vectors computed correlation was \( r \approx -0.90 \)) or positive (e.g., the \( a_1 \) and \( a_2 \) item vectors computed correlation was \( r \approx -0.90 \)). For the positively correlated \( a \)-parameters level \( (r \approx 0.9) \), first half of the items \( a \)-parameters were randomly generated from \( \text{Unif}(1.2, 1.6) \) and second half of the items \( a \)-parameters were randomly generated from. For the negatively correlated \( a \)-parameters level \( (r \approx -0.9) \), the first half of the items were set to be large \( \text{Unif}(1.2, 1.6) \) for the first dimension, and for the second dimension, \( a \)-parameters were set to be small \( \text{Unif}(0.2, .6) \). For the second half of the items, the \( a \)-parameters were set to be small \( \text{Unif}(0.2, .6) \) for the first dimension, and for the second dimension, \( a \)-parameters were set to be large \( \text{Unif}(1.2, 1.6) \). So, essentially the positively correlated case was approximate unidimensionality and the negatively correlated case was approximate simple structure. To generate the model intercept, \( d \), the multidimensional difficulty \( b \), was first randomly generated from \( \text{Norm}(0, 1) \). Once \( b \) and \( a \)-parameters were obtained for each item \( j \), the following equation was used to obtain \( d \) (Reckase, 2009): \( d_j = -\text{MDISC}_j \times b_j \) where \( \text{MDISC}_j = \sqrt{a^2_{j1} + a^2_{j2}} \) which represent the overall multidimensional discrimination of the \( j \)th item. The vector plots (Ackerman, 1996) located in Figure 1 show the item locations in a two-dimensional latent space when the \( a \)-parameters are either (a) positively or (b) negatively correlated. Once the generated 2D-MIRT model was obtained, the procedure discussed in the previous section was used to compute the linear composite scale, \( \theta_a \), along with corresponding item parameters \( a^* \) and \( d^* \). Dichotomous response data generated from the 2D-MIRT model was fitted using UIRT 2PL model to obtain the estimated unidimensional scale, \( \hat{\theta} \).

An expectation-maximization (Bock & Aitkin, 1981) algorithm was used to estimate the item parameters \( a \) and \( d \) of the UIRT model with a convergence criterion set at .0001. For the 2D-MIRT, the first item’s \( a_2 \) was fixed to zero to resolve rotational indeterminacy issues. Once the item parameter estimates were obtained, an expected a-posteriori (Embretson & Reise, 2000) estimator was used to obtain the estimated latent variable parameters, \( \hat{\theta} \). To allow for a fairer comparison between the two scales, \( \theta_a \) and \( \hat{\theta} \), both a scaling and linear transformation procedure was implemented. The scaling procedure was first implemented to standardize the \( \theta_a \) and \( \hat{\theta} \).
scales using a simple z-score conversion formula. Once \( \theta_a \) and \( \hat{\theta} \) were rescaled, a linear transformation method was implemented as follows (Kolan & Brennan, 2014)

\[
I_\theta(\theta_a) = \frac{\sigma(\hat{\theta})}{\sigma(\theta_a)} \theta_a + \left[ \mu(\hat{\theta}) - \frac{\sigma(\hat{\theta})}{\sigma(\theta_a)} \mu(\theta_a) \right]
\]

(12)

where \( \text{slope} = \frac{\sigma(\hat{\theta})}{\sigma(\theta_a)} \) and \( \text{intercept} = \mu(\hat{\theta}) - \frac{\sigma(\hat{\theta})}{\sigma(\theta_a)} \mu(\theta_a) \). A similar linear transformation procedure was done to make the results between item parameters \((a^*, \hat{a})\) and \((d^*, \hat{d})\) comparable using the same metric. Once all model parameters were rescaled, metrics such as mean absolute difference (MAD) and correlation, \( r \), were used to evaluate the results. The MAD was computed as follows

\[
MAD = \frac{\sum_{i=1}^{N} |\theta_a - \hat{\theta}|}{N}
\]

(13)

In the case of computing the MAD across item parameters \((a^*, \hat{a})\) and \((d^*, \hat{d})\), total sample size \( N \) was replaced with total test length \( J \) in the equation above. There was a total of \( 3(\text{sample size}) \times 4(\text{test length}) \times 6(\text{correlation for latent variables}) \times 2(\text{correlation for item discrimination}) = 144 \) joint conditions with each joint condition being replicated 100 times. Table S1 under supplementary materials provides a summary of the simulation design.

**Results from Simulation Experiment**

MAD results from the simulation experiment are presented in Figure 2–4. The \( r \) results from the simulation experiment are located in supplementary materials under Figures S4, S5, and S6. The values in the tables represent the average MAD and \( r \) with corresponding standard errors of MAD and \( r \) across the 100 replications within each joint condition. Results showing how similar the estimated UIRT scale \( \hat{\theta} \) is to the linear composite scale \( \theta_a \) obtained from the generated 2D-MIRT model are located in Figure 2 Results showed that on average, as the number of items increased, the MAD decreased and \( r \) results increased between \( \hat{\theta} \) and \( \theta_a \). Additionally, sample size \( N \) had minimal impact on the MAD and \( r \) between \( \hat{\theta} \) and \( \theta_a \) when \( \theta_1 \) and \( \theta_2 \) were independent or positively correlated. However, when both \( \theta_1 \) and \( \theta_2 \) and \( \theta_a \) were negatively correlated, the MAD decreased and \( r \) increased as sample size increased.

Compared to when correlation between \( \theta_1 \) and \( \theta_2 \) was negative, the MAD and \( r \) results generally improved when the correlation between \( \theta_1 \) and \( \theta_2 \) was positive, regardless of correlation between \( \theta_a \) and \( \theta_a \). When \( \theta_1 \) and \( \theta_2 \) were independent, the MAD and \( r \) values between \( \hat{\theta} \) and \( \theta_a \) were similar under the positively and negatively correlated \( \theta_1 \) and \( \theta_2 \) conditions. Figure 2 also shows that the results for \( p = -.4 \) were significantly better than for \( p = -.8 \). Indeed, the results for \( p = -.4 \) had only small differences from the results for positive ability correlation, especially for the three longer test lengths. Note that the baseline scenario for \( \theta \) from the UIRT model is when \( p = 1 \). In summary, comparing the results from the two simulation experiments for positively and negatively correlated \((\theta_1, \theta_2)\), the MAD and \( r \) results for \( \theta \) were comparable when the correlation between the latent variables \( \theta_1 \) and \( \theta_2 \) in the simulation experiment were independent or positive. In contrast, when the correlation between the latent variables \( \theta_1 \) and \( \theta_2 \) in the simulation experiment were negative, especially when \( p = -.8 \), the MAD and \( r \) results in the simulation experiment were poorer than the results from the baseline scenario.
Results showing how similar the estimated slope parameters $\hat{a}$ in the fitted UIRT model to the linear composite model $a^*$ obtained from the generated 2D-MIRT are located in Figure 3. Results showed that on average, as the sample size $N$ increased, the MAD noticeably decreased and $r$ results slightly increased between $\hat{a}$ and $a^*$. Indeed, $r$ remained close to 1 when sample size is high. Results indicated that test length had minimal impact on the MAD and $r$ between $\hat{a}$ and $a^*$ when $\theta_1$ and $\theta_2$ were orthogonal or positively correlated. For negatively correlated $a_1$ and $a_2$, MAD decreased from test length = 40 to test length = 80 but showed no further tendency to decrease with further increases in test length. When both sample size and test length were high, the MAD was low and $r$ was high even when $\theta_1$ and $\theta_2$ were highly negatively correlated (e.g., $\rho = -0.8$). The MAD result tended to improve when the correlation between $\theta_1$ and $\theta_2$ was positive. When $\rho$ was negative, the MAD and $r$ results between $\hat{a}$ and $a^*$ were better under the positively correlated $a_1$ and $a_2$ condition than the negatively correlated $a_1$ and $a_2$ condition. The baseline results for $a$ from the UIRT model is when $\rho = 1$. These results indicated that the correctly specified UIRT model ($\rho = 1$), values for $r$ for recovered $a$ - parameters were similar to the incorrectly specified UIRT model ($\rho = 0$).

Results showing how similar the estimated intercept parameters $\hat{d}$ in the fitted UIRT model to the linear composite model $d^*$ obtained from the generated 2D-MIRT are located in Figure 4. Results show that as the sample size increased, the MAD decreased and $r$ remained
close to 1 between $\hat{d}$ and $d'$. Results show that test length and correlation between both $\theta_1$ and $\theta_2$ and $a_1$ and $a_2$ had minimal impact on the MAD and $r$ results between $d$ and $d'$. The baseline results for $d$ from the UIRT model are when $\rho = 1$. These results indicated that the incorrectly specified UIRT model recovered $a$-parameters equally-well as the correctly specified UIRT model.

To further illustrate the results presented in this study, Figure S7 plots the relationship between the expected test score under UIRT and MIRT when (a) $\rho = .5$ and (b) $\rho = -.5$. The results from the plots show a positive relationship between the expected test scores when $\rho = .5$, but an independent relationship between the expected test scores when $\rho = -.5$. Figure 5 illustrates the relationship between the estimated $\theta$ from UIRT and generated $\theta_1$ and $\theta_2$ from MIRT at $\rho = .5$. The results illustrate a positive relationship between the three latent variables. Figure S4 in supplementary materials illustrates the positive relationship between estimated $\theta$ from UIRT and $\theta_0$.

**Real Data Experiment**

A real data experiment was conducted using a 20-item multiple-choice mathematics test obtained from American College Testing (ACT). The purpose of this experiment was to
close to 1 between $d$ and $d/C3$. Results show that test length and correlation between both $u1$ and $u2$ and $\alpha1$ and $\alpha2$ had minimal impact on the MAD and $r$ results between $\hat{d}$ and $d/C3$. The baseline results for $d$ from the UIRT model are when $r = 1$. These results indicated that the incorrectly specified UIRT model recovered $\alpha$-parameters equally well as the correctly specified UIRT model.

To further illustrate the results presented in this study, Figure S7 plots the relationship between the expected test score under UIRT and MIRT when (a) $r = 5$ and (b) $r = C0:5$. The results from the plots show a positive relationship between the expected test scores when $r = 5$, but an independent relationship between the expected test scores when $r = C0:5$. Figure 5 illustrates the relationship between the estimated $u$ from UIRT and generated $u1$ and $u2$ from MIRT at $r = 5$. The results illustrate a positive relationship between the three latent variables. Figure S4 in supplementary materials illustrates the positive relationship between estimated $u$ from UIRT and $ua$.

Real Data Experiment

A real data experiment was conducted using a 20-item multiple-choice mathematics test obtained from American College Testing (ACT). The purpose of this experiment was to examine how well a fitted UIRT model approximates the linear composite direction, $\thetaa$, in a multidimensional latent space when real data is presented. A sample size of 4000 was used for the analysis. The responses to the 20 multiple-choice items were dichotomously scored where 0 was coded for an incorrect response and 1 was coded for a correct response. A UIRT model and 2D-MIRT were both fitted to the observed response data. In the 2D-MIRT analysis, the 20 items were either substantively identified as “pure math” (i.e., 10 only loaded on $u1$) or “math and verbal” (i.e., 10 loaded on both $\theta1$ and $\theta2$). An expectation-maximization algorithm in the mirt package (Chalmers, 2012) was used to estimate the item parameters $a$ and $d$ for the UIRT model and $a1$, $a2$, and $d$ for the 2D-MIRT. The convergence criterion for the UIRT model and 2D-MIRT was set at .001. Once the item parameter estimates were obtained, an expected a-posteriori estimator was used to obtain the estimated latent variable parameters, ($u1$, $u2$) in the 2D-MIRT and $u$ in the UIRT model. The fitted 2D-MIRT was used to compute the linear composite direction $ua$ in equation (3) with corresponding item parameters $a^*$ and $d^*$ computed in equations (11) and (12). The estimated factor correlation between latent variables $\theta1$ and $\theta2$ was .42. After rescaling, the fitted unidimensional scale, $\hat{\theta}$ and the linear composite scale, $\hat{\thetaa}$ obtained from the fitted 2D-MIRT were compared using the MAD and $r$ metrics. Results comparing $\hat{\theta}$ to $\hat{\thetaa}$ and $d$ to $d^*$ were also examined using the MAD and $r$ metrics. Results from the
real data experiment showed that the fitted UIRT model was sufficient (i.e., acceptable MAD values) in approximating the linear composite direction based on the fitted 2D-MIRT. The MAD and $r$ between $\hat{\theta}$ and $\theta_\alpha$ was .08 and 1, respectively. Results for the slope parameters show that the MAD and $r$ between $\hat{a}$ and $\hat{a}'$ was .02 and 1, respectively. Results for the intercept parameters show that the MAD and $r$ between $\hat{d}$ and $\hat{d}'$ was .03 and 1, respectively. Item parameter estimates including $\hat{a}$ and $\hat{d}$ in the UIRT model, $\hat{a}_1$, $\hat{a}_2$, and $\hat{d}$ in the 2D-MIRT, and $\hat{a}'$ and $\hat{d}'$ in the marginal response function with respect to a linear composite are provided in Table S2 under supplementary materials. Vectors for the estimated item parameters with the corresponding linear composite direction is in Figure 6.

**Discussion**

The purpose of this study was to investigate through empirical means how well the fitted UIRT model approximates the linear composite direction $\theta_\alpha$ in a multidimensional latent space, as defined by Wang’s LCC (1987). Theoretical development on the issue is considered highly challenging, if not impossible (Junker, 1994; Junker & Stout, 1994). Our extensive simulation results overall show that the fitted UIRT model sufficiently approximates $\theta_\alpha$ when correlation between $\theta_1$ and $\theta_2$ is positive. When the correlation between $\theta_1$ and $\theta_2$ is negative, instability occurs when the fitted UIRT model is used to approximate $\theta_\alpha$. In the real data ACT example, correlation between latent traits is moderately positive and the LCC appears to be valid. Comparing results between (a) positive correlation, and (b) negative correlation between $(a_1, a_2)$, MAD and $r$ are generally slightly worse for (b). In summary, the linear approximation works best when the latent distributions are positively correlated and the discrimination parameters across dimensions are also positively correlated.

Our results suggest that the LCC is not universally true. For example, the linear composite is a rather poor approximation when the latent distributions are negatively correlated at a high level and at the same time sample size and test length are low. However, in most practical applications the correlations between abilities are positive—a phenomenon that has long been observed. Historically, Spearman (1927) used the term positive manifold to characterize different cognitive abilities that are positively correlated and attributed positive manifold to a general underlying factor, or the well-known g-factor in intelligence testing. More recent work seemed to suggest that positive manifold could also emerge purely by positive beneficial interactions between cognitive processes during development and that the underlying factor played no role.
Comparing results between (a) positive correlation, and (b) negative correlation between latent traits is moderately positive and the LCC appears to be valid. The linear composite direction is denoted by $\theta^\alpha$. Item vectors for the estimated item parameters obtained from the mathematics test dataset. The linear composite direction is denoted by $\theta^\alpha$.

Our results suggest that the LCC is not universally true. For example, the linear composite has not been thoroughly covered. One example of nonlinearity is when the mean values in $\theta_1 \theta_2$ do not follow a linear trend when conditioned on an overall measure of ability such as total score. Specific patterns of item structures may lead to such nonlinearity. Although we have investigated how different patterns of item loadings in the multidimensional space (see Figure 2), a limitation of our investigation is that the issue of possible nonlinearity composite has not been thoroughly covered. One example of nonlinearity is when the mean values in $\theta_1 \theta_2$ do not follow a linear trend when conditioned on an overall measure of ability such as total score. Specific patterns of item structures may lead to such nonlinearity. Ip et al. (2019) provided an example of such condition, which was termed non-proportional ability requirement (NPAR). Using PISA data as an example, the authors described a situation under which disproportionate cognitive demand across the dimensions is required for items that reside at certain regions (e.g., high ends) of the latent abilities. Another example is vertical scaling (Carlson, 2017; Strachan et al., in press). It is therefore possible that in some cases a nonlinear approximation of the composite will better represent the fitted UIRT. Indeed, Junker (1994) expressed skepticism about the linearity of the fitted unidimensional model for representing the multidimensional latent space (also see Kim, 1994), and Ip et al. (2013) offered a nonlinear version of the approximation, which was based on solving a system of nonlinear equations. Their preliminary results based on simulations showed that the nonlinear approximation tracked closely how multidimensional responses is represented by a single dimension, which they called the functional dimension.

The current study is also limited to the study of MIRT of two dimensions. If the number of dimension $m$ is more than two and the dimensions still form a positive manifold, then we expect the LCC continues to hold. The degree to which the approximation holds is likely also to depend on the factors we discussed above. When $m>2$ in the positive manifold, analogous to variance explained in principal component analysis, the first eigen value of $A^T A$ could also provide information of how well the fitted UIRT represents the composite direction. Another limitation of the real-data example is the use of 2PL model which ignores guessing effect. Finally, the sample size of 250 may also be too small for MIRT to function well.
By understanding the conditions under which the LCC works best, one can utilize the results of this study in different ways. Reise et al. (2014) argued that many psychological constructs have substantive breath and heterogeneous contents that result in multidimensional responses and the standard approach of using goodness-of-fit indexes to assess fit of UIRT cannot guarantee that “the common target latent ability is identified correctly or that estimated item parameters properly relate the relation between the item responses and the common latent trait.” The authors went on to suggest the approach of comparing slope parameters in UIRT and MIRT alternatives. The result in this study can be used to inform this kind of approach. For example, the LCC can be used to quickly see how parameters in a fitted UIRT change with the addition or deletion of items by examining the first eigen vector of $A^T A$, the angle of the linear composite, and the corresponding slope parameter of the fitted UIRT. The procedure can then be used to assess if an item is acceptable in terms of its deviation from the linear composite direction.

**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: This work was supported by the Institute of Education Sciences (R305D150051).

**ORCID iD**
Tyler Strachan https://orcid.org/0000-0002-3319-6332

**Supplemental Material**
Supplemental material for this article is available online.

**References**
Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*(2), 113–127. https://doi.org/10.1177/014662168901300201

Ackerman, P.L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinants. *Journal of Applied Psychology, 77*(5), 598–614. https://doi.org/10.1037/0021-9010.77.5.598

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20*(4), 311–329. https://doi.org/10.1177/014662169602000402

Ackerman, T. A., Gierl, M., & Walker, C. M (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37–51.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*, 255–278.

Ackerman, T. A., Ma, Y., Ip, E.H. (In press). Comparison of three unidimensional approaches to represent a two-dimensional latent ability space. In *Proceedings of the International Meeting of the Psychometric Society*, New York City, July 2018.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459. https://doi.org/10.1007/bf02293801
Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*(2), 129–147. https://doi.org/10.1177/014662169201600203

Carlson, J. E (2017). *Unidimensional vertical scaling in multidimensional space (ETS RR-17-29).* Educational Testing Service.

Chalmers, P. R (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Del Rosario Basterra, M., Trumbull, E., & Solano-Flores, G (2011). *Cultural validity in assessment.* Routledge.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory to multidimensional data. *Applied Psychological Measurement, 7*(2), 189–199. https://doi.org/10.1177/014662168300700207

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

Hambleton, R. K., & Swaminathan, H (1985). *Item response theory: Principles and applications.* Kluwer Nijhoff.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violation of the unidimensionality assumption. *Journal of Educational Statistics, 11*(2): 91–115.

Han, K. T., & Paek, I. (2014). A review of commercial software packages for multidimensional IRT modeling. *Applied Psychological Measurement, 38*(6), 486–498. https://doi.org/10.1177/0146621614536770

Humphreys, L.G (1982). Systematic heterogeneity of items in tests of meaningful and important psychological attributes: A rejection of unidimensionality. *Unpublished manuscript.* University of Illinois at Urbana-Champaign.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *The British Journal of Mathematical and Statistical Psychology, 63*(2), 395–416. https://doi.org/10.1348/000711009X466835

Ip, E. H., Strachan, T., Fu, Y., Chen, S., Rutkowski, L., Lay, A., Willse, J., & Ackerman, T. (2019). Bias and bias correction method for non-proportional abilities requirement (NPAR) tests. *Journal of Educational Measurement, 56*(1), 1–22. https://doi.org/10.1111/jedm.12204

Ip, E. H., Molenberghs, G., Chen, S. H., Goegebeur, Y., & De Boeck, P. (2013). Functionally Unidimensional Item Response Models for Multivariate Binary Data. *Multivariate Behavioral Research, 48*(4), 534–562. https://doi.org/10.1080/00273171.2013.796281

Junker, B. W (1994). *Inference, robustness, and essential unidimensionality.* Talk given at the 10th annual workshop on item response theory modeling. University of Twente.

Junker, B. W., & Stout, W. F (1994). Robustness of ability estimation when multiple traits are present with one trait dominant. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds), *Modern theories of measurement: Problems and issues* (pp. 31–61). University of Ottawa.

Kahraman, N., & Thompson, T. (2011). Relating unidimensional IRT parameters to a multidimensional response space: A review of two alternative projection IRT models for subscale scores. *Journal of Educational Measurement, 48*(2), 146–164. https://doi.org/10.1080/00273171.2011.100138.x

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data. *Doctoral thesis.* University of Illinois.

Kolan, M. J., & Brennan, R. L (2014). *Test equating, scaling, and linking: Methods and practices.* Springer.

Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite trait for multidimensional tests. *Applied Psychological Measurement, 16*(3), 279–293. https://doi.org/10.1177/014662169201600308

McDonald, R. P (1997). Normal-ogive multidimensional model. In R.K. Hambleton, & W. van der Linden (Ed.). *Handbook of modern item response theory* (pp. 257-269). Springer. https://doi.org/10.1007/978-1-4757-2691-6_15

R Core Team (2017). R: A language and environment for statistical computing. In: *R foundation for statistical computing.* URL. https://www.R-project.org/
Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207–230. https://doi.org/10.3102/10769986004003207

Reckase, M. D. (1985). The difficulty of test items that measure more than one dimension. *Applied Psychological Measurement, 9*(4), 401–412. https://doi.org/10.1177/014662168500900409

Reckase, M. D (2009). *Multidimensionality item response theory*. : Springer-Verlag.

Reckase, M.D., Carlson, J.E., Ackerman, T.A., & Spray, J.A (1986). The interpretation of unidimensional IRT parameters when estimated from multidimensional data. Paper presented at the annual meeting of the Psychometric Society, Toronto.

Reise, S.P., Cook, K.F., & Moore, T.M (2014). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S.P. Reise, & D.A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge. (pp. 13–40).

Spearman, C. (1927). *The abilities of man*, New York: Macmillan.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589–617. https://doi.org/10.1007/bf02294821

Strachan, T., Cho, J., Chen, S-H., Ackerman, T., Kim, K.Y., Willse, J., Weeks, J., & Ip, E.H (in press).

Using a projection IRT method for vertical scaling when construct shift is present. *Journal of Educational Measurement*.

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review, 13*, 842–60. DOI: 10.1037/0033-295X.113.4.842.

Wang, M.M. (1987). Fitting a unidimensional model to multidimensional item response data (ONR Rep. 042286). Iowa City, IA: University of Iowa.

Watson, D, Clark, LA, & Tellegen, A (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070. https://doi.org/10.1037//0022-3514.54.6.1063

Zhang, J., & Stout, W. (1999a). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 213–249. https://doi.org/10.1007/bf02294536

Zhang, J., & Stout, W. (1999b). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*(2), 129–152. https://doi.org/10.1007/bf02294532

Zhang, J., & Wang, M. (1998, April). Relating reported scores to latent traits in a multidimensional test. Paper presented at the annual meeting of the American Educational Research Association. San Diego, CA.