Unidimensional IRT Item Parameter Estimates Across Equivalent Test Forms With Confounding Specifications Within Dimensions

Educational and Psychological
Measurement

2016, Vol. 76(2) 258–279

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164415589756

epm.sagepub.com

Ki Lynn Matlock¹ and Ronna Turner²

Abstract

When constructing multiple test forms, the number of items and the total test difficulty are often equivalent. Not all test developers match the number of items and/or average item difficulty within subcontent areas. In this simulation study, six test forms were constructed having an equal number of items and average item difficulty overall. Manipulated variables were the number of items and average item difficulty within subsets of items primarily measuring one of two dimensions. Data sets were simulated at four levels of correlation (0, .3, .6, and .9). Item parameters were estimated using the Rasch and two-parameter logistic unidimensional item response theory models. Estimated discrimination and difficulty were compared across forms and to the true item parameters. The average unidimensional estimated discrimination was consistent across forms having the same correlation. Forms having a larger set of easy items measuring one dimension were estimated as being more difficult than forms having a larger set of hard items. Estimates were also investigated within subsets of items, and measures of bias were reported. This study encourages test developers to not only maintain consistent test specifications across forms as a whole but also within subcontent areas.

Keywords

item response theory (IRT), unidimensionality, multidimensional data, item difficulty, item discrimination, Rasch model

¹Oklahoma State University, Stillwater, OK, USA

²University of Arkansas, Fayetteville, AR, USA

Corresponding Author:

Ki Lynn Matlock, Research, Evaluation, Measurement, and Statistics, College of Education, Oklahoma State University, 313 Willard Hall, Stillwater, OK 74078, USA.

Email: ki.matlock@okstate.edu

Large-scale testing companies often administer multiple test forms during a given administration to a set of examinees. Each form is intended to assess the same set of content at a similar cognitive level but with a different set of items. Oftentimes, a test does not measure a single content area, but it is composed of multiple, related sub-content areas (i.e., a mathematics test composed of algebra and geometry items), creating the potential of multidimensional data. Researchers have concluded that under certain criteria, the data may still be analyzed in a unidimensional context (Reckase, 1990). However, when item difficulty is confounded within multiple dimensions, the resulting unidimensional parameter estimates may be biased (Ackerman, 1987a, 1987b, 1989; Reckase, Ackerman, & Carlson, 1988). Many test developers are careful to maintain an equal number of items and total test difficulty when constructing multiple test forms, but not all maintain an equal number of items and/or average item difficulty within subcontent areas, for example, the American College Testing (ACT, 2007), the Iowa Test of Basic Skills (ITBS; Dunbar et al., 2008), and the Oklahoma Core Curriculum Tests (OCCT; Oklahoma State Department of Education [OSDE], 2013). The purpose of this study was to investigate the effects of estimating unidimensional item parameters across test forms that have the same number of items and total test difficulty overall, yet have confounding length and difficulty within subcontent areas. Item specifications for the current simulation study were taken from the ACT Mathematics Usage Test Form 24B (Reckase & McKinley, 1991); additional forms were constructed by altering the discrimination and/or difficulty of items within sets of items primarily measuring one of two dimensions, while maintaining equal total test difficulty across all forms.

Previous studies have simulated a real testing situation where a single test is composed of subsets of items that have confounding discrimination and/or difficulty within sets, but few have compared the results across multiple test forms that are known to have equal difficulty overall. Additionally, comparisons have not been made between the unidimensional estimated discrimination and difficulty with the multidimensional *MDISC* and *MDIFF* values, respectively.

In studies of a single data set, the estimated item discrimination and difficulty were not highly biased when a test was balanced, that is, the number of items primarily measuring one dimension was equal to the number of items primarily measuring a second dimension, even when the difficulty of the of the two sets of items was different (Ackerman, 1987a). On tests having an unequal number of items within sets of items measuring only one dimension or some combination of two dimensions, the estimated item discrimination and difficulty were biased (Reckase et al., 1988). The magnitude of the bias tended to differ across studies due to the inconsistencies in true specifications. Tests with items that discriminated more heavily on one dimension than the other tended to have an estimated discrimination equal to or slightly less than the average of the true values (Ansley & Forsyth, 1985; Reckase et al., 1988; Song, 2010) or equal to the sum of the true values (Way, Ansley, & Forsyth, 1988). Tests with items equally discriminating on the two dimensions were reported to be a

lower estimate than both of the true discrimination values (Ackerman, 1987b) or a higher estimate than both of the true values (Song, 2010).

The estimated unidimensional difficulty within all sets of items was an unbiased estimate of the true multidimensional difficulty (d) in some cases (Reckase et al., 1988; Song, 2010) or an overestimate of the average of true unidimensional difficulty values (Ansley & Forsyth, 1985; Way et al., 1988). As the correlation between dimensions increased, measure of bias for item parameters tended to increase in some cases (Finch, 2010; Way et al., 1988) or decrease in others (Ansley & Forsyth, 1985; Way et al., 1988). Comparisons across studies could not be made because of inconsistent simulation specifications, models used, and software operated.

The ITBS (Dunbar et al., 2008) and the OCCT (OSDE, 2013) technical manuals indicate that the number of items within subcontent areas may differ across forms. Furthermore, few developers clearly state that difficulty for an entire test and within subcontent areas is matched across forms (Texas Education Agency [TEA], 2015a). When overall difficulty is not equal across forms, methods of equating are used to put all forms on the same scale (ACT, 2007; TEA, 2015b). Forms are equated as a whole, but not within subcontent areas, creating the potential still of subcontent areas that differ in difficulty across forms.

Current Study

In this study, six test forms were simulated; each had the same number of items and average item difficulty. Within sets of items primarily measuring one of the two dimensions, the item discrimination and/or difficulty was slightly adjusted. The goal was to create a realistic setting of multiple forms of a test that measures two content areas. For example, consider two forms of a mathematics test with items measuring both algebra and geometry. Overall, the forms had the same total number of items and average item difficulty. Form A had more algebra items at a high difficulty level and fewer easy geometry items; Form B had fewer easy algebra items and more geometry items that are hard. Even though these forms were similar overall, they differ within subcontent areas. If the same set of examinees interacted with all test forms, and the data were analyzed with the same model using the same software, how might item parameters be estimated differently across these forms? Would the average estimated total test difficulty be similar across all forms? Would the average estimated difficulty within subsets of items be affected by the confounding of the number of items and true item difficulty within dimensions? This study aims to address these questions.

A simulation study is valuable to address these research questions. By simulating data based on true parameters and analyzing the data under the same conditions, the estimated parameters can be directly compared across data sets and to the known true parameters. In practice, when test forms do not have equal average item difficulty overall, test forms are equated. However, in this simulation study, equating was unnecessary due to the construction of forms that were truly equivalent overall. This allowed for the investigation of the effects of the estimated item parameters across forms that were truly equivalent.

Unidimensional Item Response Theory

Item response theory (IRT) applies a probability model to response data in order to estimate item parameters, such as discrimination and difficulty, and to estimate examinees' ability scores. The unidimensional model assumes that the items measure a single dominant factor. The two-parameter logistic model (2PL), Equation (1), from Rizopoulos (2010), estimates the probability of a correct response to item i for an examinee with a specific ability level, θ :

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}, \quad (1)$$

where $P_i(\theta)$ is the probability of a correct response to item i from an examinee with ability θ , a_i is the discrimination parameter of item i , b_i is the difficulty parameter of item i , and θ is the ability score of the examinee. A Rasch model is the case where the discrimination factor of all items is fixed to one.

Various unidimensional IRT models are used in practice. Items on the PISA test (Kastberg, Roey, Lemanski, Chan, & Murray, 2014) and the Texas Assessment of Knowledge and Skills test (TEA, 2015a, 2015b) are evaluated using an extension of the Rasch model. Items are analyzed using a 2PL model on the Upper-Elementary Mathematics Assessment Modules administered by Educational Testing Services (Hickman, Fu, & Hill, 2012). The Rasch and the 2PL models are applied in the current study because of their common uses by large-scale testing companies both nationally and internationally.

A common algorithm for estimating item parameters is the marginal maximum likelihood estimate (MMLE) based on an assumption of the ability distribution (Bock & Leiberhan, 1970). Following the MMLE, the item parameter estimates are fixed, and the ability scores are estimated using maximum likelihood estimate or some Bayesian method (de Ayala, 2009). In this study, parameters were estimated with the methods of MMLE and the Bayesian procedure of expected a priori with the expectation-maximization and quasi-Newton algorithms.

Multidimensional Item Response Theory

Equation (2) (Reckase, 2009) represents the multidimensional IRT model where the exponent is in slope–intercept form. Here, the multidimensional item difficulty parameter takes into account the discrimination and difficulty across n dimensions:

$$P(x_{ij} = 1 | \mathbf{a}_i, d_i, \boldsymbol{\theta}_j) = \frac{e^{\mathbf{a}_i' \boldsymbol{\theta}_j + d_i}}{1 + e^{\mathbf{a}_i' \boldsymbol{\theta}_j + d_i}}, \quad (2)$$

where $P(x_{ij} = 1 | \mathbf{a}_i, d_i, \boldsymbol{\theta}_j)$ is the probability of a correct response on item i from examinee j , x_{ij} is the response to item i by examinee j (1 is correct and 0 is incorrect), $\mathbf{a}_i$ is a $n \times 1$ vector of the discrimination parameters of item i on n dimensions, d_i is a

scalar multidimensional difficulty parameter of item i , taking into account the unidimensional discrimination (a_{ik}) and difficulty (b_{ik}) across all n dimensions, such that

$$d_i = - \sum_{k=1}^n a_{ik} b_{ik}, \quad (3)$$

And θ_j is a $n \times 1$ vector of ability scores on the n dimensions for examinee j .

The maximum amount of discrimination for item i , or $MDISC_i$, is provided in Equation (4):

$$MDISC_i = \sqrt{\sum_{k=1}^n a_{ik}^2}. \quad (4)$$

The degree to which an item measures each dimension corresponds to an angle in the θ -space. This is a measure of the composite of abilities necessary to answer an item correctly. Equation (5) provides the degree to which an item measures each dimension, relative to θ_1 :

$$\alpha_i = \arccos\left(\frac{a_{i1}}{MDISC_i}\right). \quad (5)$$

For the two-dimensional case, an item with $\alpha_i = 45^\circ$ equally measures the two dimensions, and requires an equal ability in each dimension to correctly respond to item i . An item with $\alpha_i < 45^\circ$ measures the first dimension more than the second, and an item with $\alpha_i > 45^\circ$ measures the second dimension more than the first. In the current study, items are grouped into sets of items which primarily measure one dimension or that almost equally measure the two dimensions based on this measure of α_i .

The $MDIFF_i$ value, presented in Equation (6) (Reckase, 1985), takes into account the item discrimination and difficulty parameters across n dimensions:

$$MDIFF_i = - \frac{d_i}{MDISC_i}. \quad (6)$$

The interpretation of the unidimensional difficulty parameter (b_i) is the opposite of the interpretation of the multidimensional item difficulty (d_i) value, and the same as the interpretation of the $MDIFF_i$ value. An easier item, having a negative b_i value, would have a positive d_i value and a negative $MDIFF_i$ value. A more difficult item, having a positive b_i value, would have a negative d_i value and a positive $MDIFF_i$ value.

Method

Sample

Four two-dimensional true ability data sets of 1,000 examinees were created. Each followed a two-dimensional standard normal distribution and had a correlation of 0, .3, .6, or .9 between dimensions.

Data

Specifications for the first test form were taken from Form 24B of the ACT Mathematics Usage Test (Reckase & McKinley, 1991). The item specifications are provided in Table 1. The degree to which each item measured one of the two dimensions was determined by α_i (Equation 5). The items were divided into three sets where (1) items discriminated more on the first dimension, $\alpha_i < 30$, (2), items discriminated on the two dimensions somewhat equally, $30 \leq \alpha_i \leq 60$, or (3) items discriminated more on the second dimension, $\alpha_i > 60$. On the first form, 20 items were in Set 1, 11 items were in Set 2, and 9 items were in Set 3. Based on the specifications of the original ACT Form 24B, items discriminating considerably more on the first dimension (Set 1) were the easiest, as measured by the multidimensional difficulty d ($M = 0.15$, $SD = 0.57$); items discriminating on the two dimensions somewhat equally (Set 2) were moderately hard ($M = -0.43$, $SD = 0.53$); items discriminating more on the second dimension (Set 3) were very difficult ($M = -1.28$, $SD = 0.39$).

Five additional forms were constructed by altering the difficulty and/or discrimination values within sets of items. An important feature is that the original form and the additional five forms had an equal number of items and average item difficulty overall, and that the forms only varied in the number of items and/or average difficulty within the sets of items primarily measuring one of the two dimensions.

To maintain an equal total test difficulty and alter the difficulty within subsets of items on some forms, the difficulty values of the original form were first standardized. Then the values were placed at a similar location on the opposite end of the distribution (Equation 7). In this manner, items within Set 1 (which were easier) became more difficult, and items within Set 3 (which were difficult) became easier; still, the overall distribution of item difficulties remained the same.

$$d \rightarrow z_d = \frac{d - (-0.32)}{0.77} \rightarrow d' = (-1) * z_d * 0.77 + (-0.32) \quad (7)$$

The odd numbered forms (1, 3, and 5) had the original specifications of item difficulty. The even numbered forms (2, 4, and 6) had the transformed difficulty values.

The number of items within sets was also manipulated by switching the a_1 and a_2 values of selected items. Forms 1 and 2 had 20 items in Set 1 and nine items in Set 3. The discrimination values of two randomly selected items in Set 3 (Items 20 and 21) were interchanged on Forms 3 and 4 so that these items were a part of Set 1; as a result, Forms 3 and 4 had 18 and 11 items in Sets 1 and 3, respectively. Forms 5 and 6 had the same two plus three additional randomly selected items from Set 3 switched (Items 11, 19, 20, 21, and 28). Forms 5 and 6 had an almost equal number of items measuring each dimension, with the number of items being 15 and 14 in Sets 1 and 3, respectively. The number of items in Set 2 did not change across forms. All forms maintained an equal average item discrimination ($\bar{a} = 0.875$) and an equal sum of true discrimination values ($\sum_{i=1}^{40} a_{i1} + a_{i2} = 1.75$).

The summary statistics of items for the test forms as a whole and within each set of items are provided in Table 2. All six test forms as a whole had similar average

Table 1. Item Specifications of the ACT Mathematics Usage Test, Form 24B.

Item	a_{1i}	a_{2i}	$MDISC_i$	α_i	d_i	$MDIFF_i$	Set
1	1.81	0.86	2.004	25.414	1.46	-0.729	1
2	1.22	0.07	1.222	3.284	0.17	-0.139	1
3	1.57	0.36	1.611	12.915	0.67	-0.416	1
4	0.71	0.53	0.886	36.741	0.44	-0.497	2
5	0.86	0.19	0.881	12.458	0.10	-0.114	1
6	1.72	0.18	1.729	5.974	0.44	-0.254	1
7	1.86	0.29	1.882	8.862	0.38	-0.202	1
8	1.33	0.34	1.373	14.340	0.69	-0.503	1
9	1.19	1.57	1.970	52.839	0.17	-0.086	2
10	2.00	0.00	2.000	0.000	0.38	-0.190	1
11	0.87	0.00	0.870	0.000	0.03	-0.034	1
12	2.00	0.98	2.227	26.105	0.91	-0.409	1
13	1.00	0.89	1.339	41.669	-0.49	0.366	2
14	1.22	0.14	1.228	6.546	0.54	-0.440	1
15	1.27	0.47	1.354	20.308	0.29	-0.214	1
16	1.35	1.15	1.773	40.426	-0.21	0.118	2
17	1.06	0.45	1.152	23.003	0.08	-0.069	1
18	1.92	0.00	1.920	0.000	0.12	-0.063	1
19	0.96	0.22	0.985	12.907	-0.30	0.305	1
20	1.20	0.12	1.206	5.711	-0.28	0.232	1
21	1.41	0.04	1.411	1.625	-0.21	0.149	1
22	1.54	1.79	2.361	49.293	0.02	-0.008	2
23	0.54	0.23	0.587	23.070	-0.69	1.176	1
24	1.53	0.48	1.604	17.418	-0.83	0.518	1
25	0.72	0.55	0.906	37.376	-0.56	0.618	2
26	0.51	0.65	0.826	51.882	-0.49	0.593	2
27	1.66	1.72	2.390	46.017	-0.38	0.159	2
28	0.69	0.19	0.716	15.396	-0.68	0.950	1
29	0.88	1.12	1.424	51.843	-0.91	0.639	2
30	0.68	1.21	1.388	60.665	-1.08	0.778	3
31	0.24	1.14	1.165	78.111	-0.95	0.815	3
32	0.51	1.21	1.313	67.145	-1.00	0.762	3
33	0.76	0.59	0.962	37.823	-0.96	0.998	2
34	0.01	1.94	1.940	89.705	-1.92	0.990	3
35	0.39	1.77	1.812	77.574	-1.57	0.866	3
36	0.76	0.99	1.248	52.487	-1.36	1.090	2
37	0.49	1.10	1.204	65.989	-0.81	0.673	3
38	0.29	1.10	1.138	75.231	-0.99	0.870	3
39	0.48	1.00	1.109	64.359	-1.56	1.406	3
40	0.42	0.75	0.860	60.751	-1.61	1.873	3

discriminations, equal average item difficulty, and the same number of items. Within sets of items, the number of items, average discrimination, and average difficulty were adjusted. On Forms 1, 3, and 5, the first set of items primarily measuring the first dimension was easier, while the third set of items primarily measuring the second

Table 2. Number of Items, Mean, and Standard Deviation of Item Parameters Within Sets of Items.

	<i>n</i>	<i>a</i> ₁	<i>a</i> ₂	<i>α</i>	<i>d</i>
Form 1	40	1.04 (0.54)	0.71 (0.56)	34.33 (25.98)	−0.32 (0.77)
Set 1	20	1.35 (0.44)	0.28 (0.27)	11.77 (8.80)	0.16 (0.57)
Set 2	11	1.01 (0.38)	1.05 (0.47)	45.31 (6.63)	−0.43 (0.53)
Set 3	9	0.39 (0.19)	1.25 (0.37)	71.06 (9.73)	−1.28 (0.39)
Form 2	40	1.04 (0.54)	0.71 (0.56)	34.33 (25.98)	−0.32 (0.77)
Set 1	20	1.35 (0.44)	0.28 (0.27)	11.77 (8.80)	−0.81 (0.57)
Set 2	11	1.01 (0.38)	1.05 (0.47)	45.31 (6.63)	−0.22 (0.53)
Set 3	9	0.39 (0.19)	1.25 (0.37)	71.06 (9.73)	0.63 (0.39)
Form 3	40	0.98 (0.57)	0.77 (0.56)	38.46 (27.34)	−0.32 (0.77)
Set 1	18	1.36 (0.47)	0.30 (0.27)	12.67 (8.80)	0.21 (0.58)
Set 2	11	1.01 (0.38)	1.05 (0.47)	45.31 (6.63)	−0.43 (0.53)
Set 3	11	0.33 (0.21)	1.26 (0.34)	73.84 (10.71)	−1.09 (0.54)
Form 4	40	0.98 (0.57)	0.77 (0.56)	38.46 (27.34)	−0.32 (0.77)
Set 1	18	1.36 (0.47)	0.30 (0.27)	12.67 (8.80)	−0.86 (0.58)
Set 2	11	1.01 (0.38)	1.05 (0.47)	45.31 (6.63)	−0.22 (0.53)
Set 3	11	0.33 (0.21)	1.26 (0.34)	73.84 (10.71)	0.44 (0.54)
Form 5	40	0.93 (0.61)	0.82 (0.52)	43.80 (28.11)	−0.32 (0.77)
Set 1	15	1.46 (0.44)	0.34 (0.28)	13.31 (9.03)	0.31 (0.57)
Set 2	11	1.01 (0.38)	1.05 (0.47)	45.31 (6.63)	−0.43 (0.53)
Set 3	14	0.29 (0.21)	1.17 (0.35)	75.28 (10.34)	−0.92 (0.60)
Form 6	40	0.93 (0.61)	0.82 (0.52)	43.80 (28.11)	−0.32 (0.77)
Set 1	15	1.46 (0.44)	0.34 (0.28)	13.31 (9.03)	−0.96 (0.57)
Set 2	11	1.01 (0.38)	1.05 (0.47)	45.31 (6.63)	−0.22 (0.53)
Set 3	14	0.29 (0.21)	1.17 (0.35)	75.28 (10.34)	0.28 (0.60)

dimension was harder. On Forms 2, 4, and 6, the first set of items primarily measuring the first dimension was hard, and the third set of items primarily measuring the second dimension was easy. Forms 1 and 2 had a very unbalanced number of items measuring each dimension, Forms 3 and 4 were slightly unbalanced, and Forms 5 and 6 had an almost equal number of items primarily measuring each dimension.

Analysis

Parameters were estimated using the Rasch model and the 2PL model (Rizopoulos, 2010) using the “ltm” package in R (Rizopoulos, 2006). Each set of item response data (six forms across four set of examinees for a total of 24 data sets) was replicated 500 times and analyzed. For each data set, the estimated item parameters were averaged across the replications for forms as a whole and within the first, second, and third sets of items. The confounding effects of the number of items and item difficulty within dimensions were evaluated by comparing the true measures of discrimination and difficulty with the estimated unidimensional values, measured by the average bias.

According to Reckase et al. (1988), data that measure multiple dimensions may be considered unidimensional under special criteria. Prior to estimating item parameters using a unidimensional model, each data set was evaluated using the Q_3 statistic. Small negative values of Q_3 with small standard deviations indicate that the assumption of unidimensionality does hold; positive values or those with large standard deviations signify a violation. The values of the Q_3 statistic for all simulated data sets were approximately -0.025 , with a standard deviation between 0.035 and 0.080 . This confirmed that though forms were simulated using a two-dimensional model, the data could be analyzed using a unidimensional model, as is the case with many educational data sets (Kastberg et al., 2014; OSDE, 2013; TEA, 2015b; U. S. Department of Education, 2001).

Results

Discrimination

The estimated item discriminations when analyzed with the 2PL model are discussed here. Those from the Rasch model are not discussed, since the parameter was fixed to the value of one for all items. Tables 3 and 4 report the average bias of the 2PL unidimensional estimated discrimination from the true a_1 , true a_2 , average of true values, sum of true values, and $MDISC$ within sets of items and across all forms; Table 3 contains the measures of bias for the uncorrelated data set and the data set with low correlation between dimensions ($\rho = .3$), and Table 4 reports measures of bias for data sets of higher correlation ($\rho = .6$ and $\rho = .9$). Figure 1 displays the trends of the estimated bias across forms.

For forms as a whole, the average estimated discrimination was larger than the true a_1 , a_2 , and $\bar{a}$, and less than $a_1 + a_2$. When data were uncorrelated, the estimated discrimination tended to be closest to a_1 . As the correlation increased, the estimate converged closer to $MDISC$ when $\rho = .3$ or $.6$ and closer to $a_1 + a_2$ when $\rho = .9$. Across forms at the same level of correlation, the changes in item specifications within sets of items had little effect on the bias between the estimated discrimination and true measures that took into account the discrimination on both dimensions, that is, $\bar{a}$, $a_1 + a_2$, and $MDISC$. The estimated discrimination was less biased of a_1 as compared with a_2 . This was likely due to a larger number of items with $a_1 > a_2$. As forms became more balanced, the magnitude of the differences between $\hat{a}$ and a_1 or a_2 became more equal and converged to $\bar{a}$. Difficulty had little effect on all measures of bias.

The estimated discrimination of items within Set 1 (primarily measuring the first dimension) was less biased of and a small underestimate of a_1 and a large overestimate of a_2 . The measures of bias of items within Set 2 (measuring each dimension somewhat equally) was a closer estimate of $MDISC$ when $\rho \leq .3$, a close estimate of $MDISC$ and $a_1 + a_2$ when $\rho = .6$, and a closer estimate of $a_1 + a_2$ when $\rho = .9$; the measures of bias of items within Set 2 were almost equal between $\hat{a}$ and a_1 , a_2 , and $\bar{a}$. For those items in Set 3 (primarily measuring the second dimension), the estimate tended to be closer to $\bar{a}$ when $\rho = 0$ and a_2 or $MDISC$ when $\rho > 0$. An interesting

Table 3. Average Bias of the Item Discrimination Using the Two-Parameter Logistic Model ($\rho = 0$ and $\rho = .3$).

		$\rho = 0$										$\rho = .3$									
n		$\sum \frac{(a_1 - \bar{a})}{k}$	$\sum \frac{(a_2 - \bar{a})}{k}$	$\sum \frac{(\bar{a} - \bar{a})}{k}$	$\sum \frac{(\sum \bar{a} - \bar{a})}{k}$	$\sum \frac{(MDISC - \bar{a})}{k}$	$\sum \frac{(a_1 - \bar{a})}{k}$	$\sum \frac{(a_2 - \bar{a})}{k}$	$\sum \frac{(\bar{a} - \bar{a})}{k}$	$\sum \frac{(\sum \bar{a} - \bar{a})}{k}$	$\sum \frac{(MDISC - \bar{a})}{k}$	$\sum \frac{(a_1 - \bar{a})}{k}$	$\sum \frac{(a_2 - \bar{a})}{k}$	$\sum \frac{(\bar{a} - \bar{a})}{k}$	$\sum \frac{(\sum \bar{a} - \bar{a})}{k}$	$\sum \frac{(MDISC - \bar{a})}{k}$					
Form 1	40	-0.116	-0.447	-0.281	0.594	0.243	-0.272	-0.603	-0.437	0.438	0.087	-0.272	-0.603	-0.437	0.438	0.087					
Set 1	20	0.122	-0.949	-0.414	0.403	0.168	0.057	-1.015	-0.479	0.337	0.103	0.057	-1.015	-0.479	0.337	0.103					
Set 2	11	-0.293	-0.250	-0.272	0.757	0.162	-0.530	-0.487	-0.509	0.520	-0.075	-0.530	-0.487	-0.509	0.520	-0.075					
Set 3	9	-0.428	0.429	0.001	0.819	0.508	-0.685	0.172	-0.257	0.562	0.250	-0.685	0.172	-0.257	0.562	0.250					
Form 2	40	-0.115	-0.447	-0.281	0.594	0.243	-0.292	-0.624	-0.458	0.417	0.066	-0.292	-0.624	-0.458	0.417	0.066					
Set 1	20	0.149	-0.923	-0.387	0.429	0.195	0.054	-1.018	-0.482	0.334	0.100	0.054	-1.018	-0.482	0.334	0.100					
Set 2	11	-0.321	-0.278	-0.299	0.729	0.134	-0.557	-0.514	-0.535	0.493	-0.102	-0.557	-0.514	-0.535	0.493	-0.102					
Set 3	9	-0.452	0.404	-0.024	0.794	0.483	-0.738	0.118	-0.310	0.508	0.197	-0.738	0.118	-0.310	0.508	0.197					
Form 3	40	-0.156	-0.365	-0.260	0.615	0.264	-0.324	-0.533	-0.428	0.447	0.096	-0.324	-0.533	-0.428	0.447	0.096					
Set 1	18	0.166	-0.888	-0.361	0.469	0.217	0.074	-0.980	-0.453	0.377	0.125	0.074	-0.980	-0.453	0.377	0.125					
Set 2	11	-0.338	-0.296	-0.317	0.712	0.117	-0.560	-0.517	-0.538	0.490	-0.105	-0.560	-0.517	-0.538	0.490	-0.105					
Set 3	11	-0.501	0.422	-0.040	0.756	0.487	-0.740	0.183	-0.279	0.517	0.248	-0.740	0.183	-0.279	0.517	0.248					
Form 4	40	-0.159	-0.368	-0.263	0.612	0.261	-0.345	-0.554	-0.450	0.426	0.075	-0.345	-0.554	-0.450	0.426	0.075					
Set 1	18	0.195	-0.860	-0.332	0.498	0.246	0.070	-0.985	-0.457	0.373	0.121	0.070	-0.985	-0.457	0.373	0.121					
Set 2	11	-0.361	-0.319	-0.340	0.689	0.094	-0.583	-0.540	-0.562	0.467	-0.128	-0.583	-0.540	-0.562	0.467	-0.128					
Set 3	11	-0.535	0.388	-0.074	0.722	0.453	-0.786	0.137	-0.324	0.471	0.202	-0.786	0.137	-0.324	0.471	0.202					
Form 5	40	-0.200	-0.303	-0.252	0.623	0.273	-0.374	-0.477	-0.426	0.450	0.099	-0.374	-0.477	-0.426	0.450	0.099					
Set 1	18	0.231	-0.894	-0.332	0.567	0.288	0.105	-1.020	-0.458	0.441	0.162	0.105	-1.020	-0.458	0.441	0.162					
Set 2	11	-0.373	-0.330	-0.352	0.677	0.082	-0.578	-0.536	-0.557	0.472	-0.123	-0.578	-0.536	-0.557	0.472	-0.123					
Set 3	11	-0.526	0.351	-0.088	0.642	0.405	-0.726	0.150	-0.288	0.442	0.205	-0.726	0.150	-0.288	0.442	0.205					
Form 6	40	-0.204	-0.307	-0.255	0.620	0.269	-0.395	-0.498	-0.446	0.429	0.078	-0.395	-0.498	-0.446	0.429	0.078					
Set 1	18	0.266	-0.859	-0.296	0.602	0.324	0.098	-1.026	-0.464	0.434	0.156	0.098	-1.026	-0.464	0.434	0.156					
Set 2	11	-0.392	-0.349	-0.371	0.658	0.063	-0.597	-0.554	-0.575	0.453	-0.142	-0.597	-0.554	-0.575	0.453	-0.142					
Set 3	11	-0.559	0.318	-0.121	0.609	0.372	-0.764	0.112	-0.326	0.404	0.167	-0.764	0.112	-0.326	0.404	0.167					

Table 4. Average Bias of the Item Discrimination Using the Two-Parameter Logistic Model ($\rho = .6$ and $\rho = .9$).

		$\rho = .6$					$\rho = .9$				
n		$\frac{\sum (a_1 - \bar{a})}{k}$	$\frac{\sum (a_2 - \bar{a})}{k}$	$\frac{\sum (\bar{a} - \bar{a})}{k}$	$\frac{\sum (\sum a - \bar{a})}{k}$	$\frac{\sum (MDISC - \bar{a})}{k}$	$\frac{\sum (a_1 - \bar{a})}{k}$	$\frac{\sum (a_2 - \bar{a})}{k}$	$\frac{\sum (\bar{a} - \bar{a})}{k}$	$\frac{\sum (\sum a - \bar{a})}{k}$	$\frac{\sum (MDISC - \bar{a})}{k}$
Form 1	40	-0.460	-0.791	-0.626	0.249	-0.101	-0.566	-0.897	-0.732	0.143	-0.208
Set 1	20	-0.099	-1.170	-0.635	0.182	-0.053	-0.168	-1.240	-0.704	0.112	-0.122
Set 2	11	-0.753	-0.710	-0.731	0.297	-0.297	-0.864	-0.821	-0.842	0.186	-0.408
Set 3	9	-0.905	-0.048	-0.477	0.342	0.030	-1.088	-0.231	-0.659	0.159	-0.152
Form 2	40	-0.463	-0.794	-0.628	0.247	-0.104	-0.567	-0.898	-0.732	0.143	-0.208
Set 1	20	-0.089	-1.161	-0.625	0.191	-0.043	-0.167	-1.239	-0.703	0.113	-0.121
Set 2	11	-0.759	-0.716	-0.737	0.291	-0.304	-0.870	-0.828	-0.849	0.180	-0.415
Set 3	9	-0.930	-0.073	-0.502	0.317	0.005	-1.083	-0.226	-0.655	0.164	-0.147
Form 3	40	-0.515	-0.724	-0.620	0.256	-0.095	-0.626	-0.835	-0.730	0.145	-0.206
Set 1	18	-0.099	-1.153	-0.626	0.204	-0.048	-0.184	-1.238	-0.711	0.119	-0.133
Set 2	11	-0.774	-0.731	-0.752	0.276	-0.318	-0.873	-0.830	-0.851	0.177	-0.418
Set 3	11	-0.938	-0.015	-0.476	0.319	0.051	-1.103	-0.179	-0.641	0.155	-0.114
Form 4	40	-0.518	-0.727	-0.623	0.252	-0.099	-0.627	-0.836	-0.731	0.144	-0.207
Set 1	18	-0.092	-1.146	-0.619	0.211	-0.041	-0.182	-1.237	-0.710	0.120	-0.132
Set 2	11	-0.771	-0.729	-0.750	0.279	-0.316	-0.877	-0.834	-0.856	0.173	-0.422
Set 3	11	-0.963	-0.040	-0.502	0.294	0.025	-1.105	-0.181	-0.643	0.152	-0.116
Form 5	40	-0.566	-0.669	-0.618	0.257	-0.094	-0.677	-0.780	-0.728	0.147	-0.204
Set 1	18	-0.095	-1.219	-0.657	0.241	-0.037	-0.197	-1.322	-0.760	0.139	-0.140
Set 2	11	-0.781	-0.738	-0.759	0.269	-0.326	-0.878	-0.835	-0.856	0.172	-0.423
Set 3	11	-0.903	-0.026	-0.465	0.265	0.028	-1.033	-0.156	-0.594	0.135	-0.101
Form 6	40	-0.567	-0.670	-0.619	0.256	-0.094	-0.676	-0.780	-0.728	0.147	-0.204
Set 1	18	-0.087	-1.212	-0.650	0.249	-0.030	-0.197	-1.322	-0.760	0.139	-0.140
Set 2	11	-0.780	-0.737	-0.759	0.270	-0.325	-0.878	-0.836	-0.857	0.172	-0.423
Set 3	11	-0.914	-0.037	-0.475	0.254	0.018	-1.031	-0.155	-0.593	0.137	-0.100

Figure 1. Line graph of the average bias between true and estimated measures of discrimination across six forms at four levels of correlation. A value greater than zero indicates that the true value was underestimated; a value less than zero indicates that the true value was overestimated.

pattern was presented; the estimated bias between $\hat{a}$ and *MDISC* tended to follow the same trend as the bias between $\hat{a}$ and a_1 for the first set of items and $\hat{a}$ and a_2 for the third set of items. This was likely because *MDISC* was a measure of the composite discrimination (see Equation 4). Therefore, within sets of items primarily measuring one of two dimensions, the estimated discrimination was a more consistent estimate of *MDISC* than either true unidimensional parameters; this measure of bias decreased as correlation increased.

Across forms at the same level of correlation, as forms became more balanced, the measure of bias between $\hat{a}$ and a_1 increased in magnitude, and the measure of bias between $\hat{a}$ and a_2 decreased. On forms having an almost equal number of items within sets (Forms 5 and 6), the bias between $\hat{a}$ and a_1 for items in Set 1 was almost equal to the bias between $\hat{a}$ and a_2 for items in Set 2. The changes in difficulty of subsets of items across forms had little effect on the estimated discrimination.

Difficulty

Overall the 24 data sets had the same true average difficulty (d). Due to the opposing interpretations of the multidimensional d difficulty and the unidimensional b difficulty, where $d > 0$ indicates an easier item and $d < 0$ indicates a harder item, and where $b > 0$ indicates a harder item and $b < 0$ indicates an easier item, the average bias compared the estimated $\hat{b}$ with $-d$. It should be noted that overall, measures of bias for the difficulty parameter were much smaller than those of the discrimination parameter. Table 5 reports measures of bias across all forms between the true difficulties ($-d$ and *MDIFF*) and the estimated difficulty when the Rasch model was applied; Table 6 reports this when the 2PL model was applied. Figure 2 displays these measures of bias on a line graph.

Rasch Model Results. Changes in the item difficulty within subsets of items primarily measuring one of two dimensions, even when the overall average difficulty across forms was equal, resulted in test forms with different estimated total test difficulty. Across forms with similar levels of correlation, the range of estimated difficulty was approximately 0.08. Odd-numbered forms, that is, those with a larger set of easy items and smaller set of hard items, were estimated as being more difficult, overall, than even-numbered forms, that is, those with a larger set of hard items and a smaller set of easy items. Forms with a larger set of easy items had an almost equal measure of average bias between $\hat{b}$ and either $-d$ or *MDIFF* due to the almost equal values of $-d$ and *MDIFF* on odd forms; the magnitude of bias tended to vary across levels of correlation, but with no pattern. Forms with a larger set of hard items and smaller set of easy items, had a positive bias between $\hat{b}$ and $-d$ (at all levels of correlation except $\rho = .6$ where the measure of bias was almost zero) indicating an underestimate of the true difficulty, and a negative bias between $\hat{b}$ and *MDIFF*, indicating an overestimate of the true difficulty. Overall, the estimates of difficulty were not affected by changes in the numbers of items within subsets of items.

Table 5. Average Bias of the Item Difficulty Using the Rasch Model.

		$\rho = 0$		$\rho = .3$		$\rho = .6$		$\rho = .9$	
	n	$\frac{\sum_k ((-d) - \hat{b})}{k}$	$\frac{\sum_k ((MDIFF) - \hat{b})}{k}$	$\frac{\sum_k ((-d) - \hat{b})}{k}$	$\frac{\sum_k ((MDIFF) - \hat{b})}{k}$	$\frac{\sum_k ((-d) - \hat{b})}{k}$	$\frac{\sum_k ((MDIFF) - \hat{b})}{k}$	$\frac{\sum_k ((-d) - \hat{b})}{k}$	$\frac{\sum_k ((MDIFF) - \hat{b})}{k}$
Form 1	40	-0.050	-0.060	0.016	0.007	-0.093	-0.102	-0.003	-0.012
	Set 1	-0.101	0.040	-0.043	0.098	-0.134	0.007	-0.042	0.100
	Set 2	-0.035	-0.103	0.052	-0.015	-0.061	-0.129	0.039	-0.028
Set 3	11	0.044	-0.229	0.103	-0.170	-0.040	-0.313	0.032	-0.241
Form 2	40	0.026	-0.094	0.088	-0.031	-0.009	-0.128	0.078	-0.041
	Set 1	0.080	-0.183	0.117	-0.146	0.021	-0.242	0.096	-0.167
	Set 2	11	0.007	-0.064	0.092	0.021	-0.009	-0.080	0.085
Set 3	9	-0.073	0.069	0.019	0.161	-0.074	0.068	0.029	0.171
Form 3	40	-0.045	-0.054	0.019	0.009	-0.088	-0.097	-0.002	-0.011
	Set 1	-0.108	0.055	-0.049	0.113	-0.139	0.024	-0.046	0.117
	Set 2	11	-0.030	-0.097	0.053	-0.014	-0.054	-0.122	0.037
Set 3	11	0.045	-0.189	0.095	-0.138	-0.039	-0.272	0.030	-0.203
Form 4	40	0.024	-0.095	0.089	-0.031	-0.009	-0.128	0.077	-0.042
	Set 1	0.091	-0.191	0.127	-0.155	0.033	-0.249	0.105	-0.177
	Set 2	11	0.004	-0.067	0.091	0.020	-0.012	-0.084	0.084
Set 3	11	-0.065	0.034	0.022	0.121	-0.075	0.024	0.024	0.123
Form 5	40	-0.042	-0.051	0.022	0.013	-0.084	-0.094	-0.001	-0.010
	Set 1	-0.116	0.061	-0.053	0.124	-0.140	0.038	-0.043	0.134
	Set 2	11	-0.027	-0.094	0.058	-0.009	-0.053	-0.120	0.038
Set 3	14	0.027	-0.137	0.074	-0.090	-0.050	-0.214	0.014	-0.150
Form 6	40	0.024	-0.095	0.086	-0.033	-0.010	-0.129	0.078	-0.042
	Set 1	0.116	-0.228	0.153	-0.191	0.060	-0.285	0.135	-0.210
	Set 2	11	0.003	-0.069	0.085	0.014	-0.013	-0.084	0.086
Set 3	14	-0.058	0.027	0.014	0.098	-0.081	0.003	0.011	0.095

Table 6. Average Bias of the Item Difficulty Using the two-parameter logistic Model.

	n	$\rho = 0$			$\rho = .3$			$\rho = .6$			$\rho = .9$		
		$\frac{\sum((-d)-\hat{b})}{k}$	$\frac{\sum(MDIFF)-\hat{b}}{k}$	$\frac{\sum((-d)-\hat{b})}{k}$	$\frac{\sum(MDIFF)-\hat{b}}{k}$	$\frac{\sum((-d)-\hat{b})}{k}$	$\frac{\sum(MDIFF)-\hat{b}}{k}$	$\frac{\sum((-d)-\hat{b})}{k}$	$\frac{\sum(MDIFF)-\hat{b}}{k}$	$\frac{\sum((-d)-\hat{b})}{k}$	$\frac{\sum(MDIFF)-\hat{b}}{k}$	$\frac{\sum((-d)-\hat{b})}{k}$	$\frac{\sum(MDIFF)-\hat{b}}{k}$
Form 1	40	-0.125	-0.135	0.021	0.012	-0.006	-0.016	0.076	0.067	0.076	0.067	0.076	0.067
Set 1	20	-0.174	-0.033	-0.119	0.022	-0.185	-0.044	-0.125	0.016	-0.125	0.016	-0.125	0.016
Set 2	11	0.000	-0.067	0.125	0.057	0.088	0.021	0.166	0.099	0.166	0.099	0.166	0.099
Set 3	9	-0.170	-0.443	0.207	-0.066	0.275	0.002	0.414	0.141	0.414	0.141	0.414	0.141
Form 2	40	0.096	-0.024	0.144	0.025	0.084	-0.035	0.146	0.027	0.146	0.027	0.146	0.027
Set 1	20	0.172	-0.091	0.257	-0.006	0.238	-0.025	0.314	0.051	0.314	0.051	0.314	0.051
Set 2	11	0.028	-0.044	0.112	0.040	0.050	-0.021	0.114	0.042	0.114	0.042	0.114	0.042
Set 3	9	0.009	0.151	-0.069	0.073	-0.218	-0.076	-0.188	-0.046	-0.188	-0.046	-0.188	-0.046
Form 3	40	-0.111	-0.120	0.026	0.017	-0.008	-0.018	0.086	0.077	0.086	0.077	0.086	0.077
Set 1	18	-0.194	-0.031	-0.138	0.025	-0.211	-0.048	-0.136	0.027	-0.136	0.027	-0.136	0.027
Set 2	11	0.004	-0.063	0.125	0.057	0.087	0.020	0.174	0.106	0.174	0.106	0.174	0.106
Set 3	11	-0.091	-0.324	0.196	-0.037	0.228	-0.005	0.362	0.129	0.362	0.129	0.362	0.129
Form 4	40	0.068	-0.052	0.134	0.015	0.077	-0.042	0.148	0.028	0.148	0.028	0.148	0.028
Set 1	18	0.167	-0.115	0.267	-0.015	0.256	-0.026	0.340	0.058	0.340	0.058	0.340	0.058
Set 2	11	0.027	-0.044	0.111	0.040	0.046	-0.025	0.117	0.045	0.117	0.045	0.117	0.045
Set 3	11	-0.054	0.045	-0.059	0.040	-0.185	-0.086	-0.136	-0.037	-0.136	-0.037	-0.136	-0.037
Form 5	40	-0.101	-0.111	0.030	0.020	-0.008	-0.017	0.083	0.074	0.083	0.074	0.083	0.074
Set 1	15	-0.196	-0.019	-0.150	0.027	-0.232	-0.055	-0.164	0.014	-0.164	0.014	-0.164	0.014
Set 2	11	0.008	-0.059	0.127	0.060	0.084	0.016	0.171	0.104	0.171	0.104	0.171	0.104
Set 3	14	-0.086	-0.250	0.146	-0.018	0.160	-0.004	0.279	0.115	0.279	0.115	0.279	0.115
Form 6	40	0.048	-0.071	0.127	0.008	0.072	-0.047	0.148	0.029	0.148	0.029	0.148	0.029
Set 1	15	0.209	-0.135	0.328	-0.017	0.324	-0.021	0.413	0.068	0.413	0.068	0.413	0.068
Set 2	11	0.029	-0.042	0.108	0.037	0.046	-0.025	0.119	0.048	0.119	0.048	0.119	0.048
Set 3	14	-0.109	-0.024	-0.074	0.011	-0.176	-0.092	-0.113	-0.029	-0.113	-0.029	-0.113	-0.029

Figure 2. Line graph of the average bias between true and estimated measures of difficulty across six forms at four levels of correlation.

Within sets of items that varied in numbers of items and average item difficulty within subsets, the difficulty of the harder set of items was consistently an underestimate of $-d$ and an overestimate of $MDIFF$, while the difficulty of the easier set of items was an overestimate of $-d$ and an underestimate of $MDIFF$. Furthermore, values were overestimated to a larger degree than they were underestimated. This gives light to the differences in the estimated difficulty for forms as a whole. The odd-numbered forms had a larger set of easy items with a difficulty that was an overestimate of $-d$ and a smaller set of hard items with a difficulty that was an underestimate of $-d$; overall, these were estimated as being more difficult. The even-numbered forms had a larger set of hard items with a difficulty that was an underestimate of $-d$ and a smaller set of easy items with a difficulty that was an overestimate of $-d$; overall, these were estimated as being less difficult. The bias in difficulty resulting from the larger set of items had a larger impact on the overall estimated difficulty.

As the level of correlation increased, the estimated difficulty of the smaller set of items (Set 3) became more stable across forms as the same level of correlation, and the measures of bias between $\hat{b}$ and $-d$ tended to converge to zero. The estimated difficulty of the larger set of items still tended to vary across forms, as did the total test difficulty.

Two-Parameter Logistic Model Results. The trends of the estimated difficulty were inconsistent across forms at the four levels of correlation; the ranges of estimated difficulty when the 2PL model was used tended to decrease as correlation increased. Similar to the results when the Rasch model was applied, on forms with a larger set of easy items and smaller set of hard items, the average bias between $\hat{b}$ and $-d$ or $MDIFF$ was almost equal. On forms with a larger set of hard items and smaller set of easy items, the bias between $\hat{b}$ and $-d$ was larger than that between $\hat{b}$ and $MDIFF$.

The estimated difficulty of the larger set of items tended to follow a similar pattern when estimated with either the Rasch or the 2PL model. The difficulty of these items was an overestimate of $-d$ on the odd forms (when true difficulty of this set was easy) and an underestimate on the even forms (when true difficulty of this set was hard). The smaller set of items did not follow a similar trend when data were uncorrelated; in this case, $\hat{b}$ was almost always an overestimate of $-d$, but the measure of bias was always smaller than that of the larger set of items on corresponding forms. At all other levels of correlation ($\rho \geq .3$), the difficulty of the smaller set of items was underestimated on odd forms (when true difficulty of this set was hard) and overestimated on even forms (when true difficulty of this set was easy). This was a similar trend to that of the smaller set of items when estimated with the Rasch model; however, as the correlation increased, the measure of variability across forms also tended to increase.

Comparison of the Rasch and 2PL Estimates of Difficulty. Figure 3 displays the true difficulty ($-d$), $MDIFF$, and the estimated difficulty from the Rasch and 2PL models. The true difficulty across all forms remained constant, while $MDIFF$ took into account the changes of item discriminations. The estimated total test difficulty was affected by the changes in difficulty within subsets of items, the degree of correlation

Figure 3. Line graph of the true difficulty ($-d$ and MDIFF) and the estimated difficulty when the Rasch and two-parameter logistic models were used across all forms and at all levels of correlation.

between dimensions, and the model applied. When either model was applied, forms with a larger number of easy items were estimated as being more difficult than forms with a larger number of hard items, even when overall test difficulty was equal. This trend followed the overall pattern of *MDIFF*.

The changes in the difficulty of subsets of items had a strong effect on the 2PL estimates when $\rho = 0$, yet the effect weakened as correlation increased. When the Rasch model was applied, the effects of the differences in subset difficulty remained constant at all levels of correlation. Across forms at the same level of correlation, the estimated difficulty from the Rasch model was most often a better estimate of the true difficulty ($-d$).

In conclusion, the estimated difficulty of the Rasch model tended to be closer to the true difficulty ($-d$) more often than the estimated difficulty from the 2PL model. However, the estimated difficulty from the 2PL model tended to follow the pattern of *MDIFF* more closely than did the estimated difficulty of the Rasch model.

Discussion

This study investigated one of many issues when applying a unidimensional IRT model to data that is composed of multiple dimensions—the effects of confounding item specifications within dimensions across multiple test forms that have equivalent specifications overall. The specifications of Form 24B of the ACT Mathematics Usage Test (Reckase & McKinley, 1991) were used as Form 1; five additional forms were constructed with a slightly altered number of items and/or average item difficulty within sets of items primarily measuring one of the two dimensions. As a result, the average unidimensional estimated discrimination was affected more by the correlation between dimensions, but not the confounding within dimensions. The average unidimensional estimated difficulty was affected by the confounding of specifications within dimensions, the correlation, and the model applied (Rasch or 2PL IRT models).

The estimated discrimination and difficulty of items are used by testing companies for various purposes. When constructing forms, items are evaluated and selected based on the estimated parameters. After administration, forms are equated based on the average estimated difficulty of all items across forms. Results of this study demonstrate that the average estimated difficulty of items is likely to vary across forms as a whole and within sets of items primarily measuring a single dimension when forms differ in the number of items and/or item difficulty within subsets, even when the true difficulty is equal across forms. Forms having a larger set of easy items measuring one dimension and a smaller set of hard items measuring a second dimension are likely to be estimated with a higher average difficulty overall than forms having a larger set of hard items measuring one dimension and a small set of easy items measuring another dimension. The unidimensional difficulty tends to be a close estimate of *MDIFF* when the 2PL model is used, and the estimate becomes more stable as the correlation increases. However, the estimate tends to be closer to

— d more often across forms when the Rasch model is used, yet the stability of the estimate was similar at all levels of correlation.

The estimated average item discrimination across forms (at the same level of correlation) tends to be consistent regardless of changes in difficulty and/or numbers of items within sets. When data are uncorrelated, the estimated discrimination may be a closer approximation to the average of true values, as reported by Ansley and Forsyth (1985), Reckase et al. (1988), and Song (2010), or to the discrimination of the larger set of items. As correlation increases ($\rho = .3$ and $.6$), the estimate tends to be closer to *MDISC*. When data are highly correlated ($\rho = .9$), the estimated discrimination is a close estimate of the sum of true values, which was reported by Way et al. (1988). Comparing these results with those of previous studies should be done with caution due to differences in specifications and models.

The use of simulations is valuable to study this situation. Bolt (1999) argued for the advantages of using simulations to study data with known multidimensional structures. It allows for comparisons of estimated item parameters with the true values, which are never known in practice. Forms that really do have equal true difficulty values may be estimated as having differing difficulties due to inconsistencies of properties within subsets of items. In practice, if the multiple forms were estimated with different average item difficulty, then they would most likely be equated. However, in this case forms should not be equated because they were known to have equal average item difficulties. The results call attention to the need for strict test specifications within subcontents. Additionally, the correlation between dimensions should be well determined.

Significance and Future Study

In situations where multiple test forms are administered, test developers may want to more closely investigate the item specifications within subcontent areas, in order to construct test forms that yield consistent item parameter estimates. The results of this study may be used to guide test developers to design forms with more consistent item difficulty within subsets of items. It is also advised to evaluate the correlation between the multiple dimensions. Once the correlation has been established, the unidimensional estimate from a Rasch model as compared with a 2PL model may represent the true parameters somewhat differently. If data are highly correlated, it may be desired to use a 2PL model over a Rasch model due to the stability of the estimated total test difficulty. The inclusion of the 3PL model is of interest in future studies.

Future research should focus on the study of effects on estimated ability, processes of equating, and implications in the computerized-adaptive testing setting. In some contexts, latent ability estimates are conditional on the estimated item discrimination and difficulty. Biased item parameters are likely to lead to biased ability estimates. An extension of this study is to investigate the effects on estimated ability.

After the administration of tests, procedures of equating are used to align the estimated average item difficulty across forms. This is often applied to forms as a whole,

but not within subcontent areas. Future studies should examine the effects of confounding difficulty within dimensions on equating for tests as a whole and within subcontent areas. Furthermore, this study may have implications in an adaptive setting when various methods of item selection, content balance, and termination are implemented using IRT techniques.

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) received no financial support for the research, authorship, and/or publication of this article.

References

- Ackerman, T. A. (1987a, September). *A comparison study of the unidimensional IRT estimation of compensatory and noncompensatory multidimensional item response data* (ACT Research Report Series No. 87-12). Retrieved from http://www.act.org/research/researchers/reports/pdf/ACT_RR87-12.pdf
- Ackerman, T. A. (1987b, September). *The use of unidimensional item parameter estimates of multidimensional items in adaptive testing* (ACT Research Report Series No. 87-13). Retrieved from http://www.act.org/research/researchers/reports/pdf/ACT_RR87-13.pdf
- Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, 13, 113-127.
- Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (NCES 2001-509). Washington, DC: National Center for Education Statistics, U.S. Department of Education.
- American College Testing Program. (2007). *The ACT technical manual*. Iowa City, IA: Author.
- Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9, 37-48.
- Bock, R. D., & Leiberhan, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179-198.
- Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education*, 12, 383-407.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.
- Dunbar, S. B., Hoover, H. D., Frisbie, D. A., Oberley, K. R., Bray, G. B., Naylor, R. J., . . . Hazen, T. (2008). *The Iowa tests interpretive guide for school administrators*. Rolling Meadows, IL: Riverside.
- Finch, H. (2010). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement*, 35, 67-82.

- Hickman, J., Fu, J., & Hill, H. (2012). *Technical report: Creation and dissemination of upper-elementary mathematics assessment modules* (National Science Foundation Math and Science Partnership Program (NSF 08-525) 2009-2012). Arlington, VA: The National Science Foundation.
- Kastberg, D., Roey, S., Lemanski, N., Chan, J. Y., & Murray, G. (2014, April). *2012 Data files and database with U.S.-specific variables* (Technical report and user guide for the Program for International Student Assessment (PISA) (NCES 2014-025)). Washington, DC: NCES, IES, U.S. Department of Education.
- Oklahoma State Department of Education. (2013). Test and item specifications mathematics grade 5. Retrieved from http://ok.gov/sde/sites/ok.gov.sde/files/OCCT_G5M_ItemSpecs_2012-13.pdf
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401-412.
- Reckase, M. D. (1990, April 16-20). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests*. Paper presented at the *Annual Meeting of the American Educational Research Association*, Boston, MA.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, 25, 193-203.
- Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, 15, 361-373.
- Rizopoulos, D. (2006). Ltm: An R package for latent variable modelling and item response theory analysis. *Journal of Statistical Software*, 17(5), 1-25.
- Rizopoulos, D. (2010, January 12). *Item response theory in R using package ltm* [PowerPoint slides]. Retrieved from <http://statmath.wu.ac.at/research/talks/resources/PresIRT.pdf>
- Song, T. (2010). *The effect of fitting a unidimensional IRT model to multidimensional data in content-balanced computer adaptive testing* (Doctoral dissertation). Retrieved from ProQuest database. (UMI No. 3435117)
- Texas Education Agency. (2015a). Building a high-quality assessment system. *Technical Digest 2013–2014* (Chapter 2). Retrieved from http://tea.texas.gov/Student_Testing_and_Accountability/Testing/Student_Assessment_Overview/Technical_Digest_2013-2014/
- Texas Education Agency. (2015b). Standard technical processes. *Technical Digest 2013–2014* (Chapter 3). Retrieved from http://tea.texas.gov/Student_Testing_and_Accountability/Testing/Student_Assessment_Overview/Technical_Digest_2013-2014/
- U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. (2001). *The NAEP 1998 Technical Report*, NCES 2001-509, by Allen, N. L., Donoghue, J. R., & Schoeps, T. L. Washington, DC: National Center for Education Statistics.
- Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement*, 12, 239-252.