Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 28 Number 9, June 2023 ISSN 1531-7714

Application of Two-Parameter Item Response Theory for Determining Form-Dependent Items on Exams Using Different Item Orders

Thomas C. Pentecost, *Grand Valley State University*,
Jeffery R. Raker, *University of South Florida*,
Kristen L. Murphy, *University of Wisconsin - Milwaukee*

Using multiple versions of an assessment has the potential to introduce item environment effects. These types of effects result in version dependent item characteristics (i.e., difficulty and discrimination). Methods to detect such effects and resulting implications are important for all levels of assessment where multiple forms of an assessment are created. This report describes a novel method for identifying items that do and do not display form dependence. The first two steps identify form dependent items using a differential item functioning (DIF) analysis of item parameters estimated by Item Response Theory. The method is illustrated using items that appeared in four forms (two trial and two released versions) of a first semester general chemistry examination. Eighteen of fifty-six items were identified as having item parameters that were form dependent. Thirteen of those items displayed a form dependence consistent with reasons previously identified in the literature: preceding item difficulty, content priming, and a combination of preceding item difficulty and content priming. The remaining five items had form dependence that did not align reasons reported in the literature. An analysis was done to determine if all possible instances of predicted form dependence could be found. Several instances where form dependence could have been found, based on the preceding item difficulty or content priming, were identified, and those items did not display form dependence. We identify and rationalize form dependence for thirteen of the eighteen items flagged; however, we are unable to predict form dependence for items.

Keywords: assessment, item-order effect

Introduction

The construction of any test or assessment includes several considerations for item properties. Commonly, these include internal item properties (*e.g.*, content, format) and external item properties (*e.g.*, item-order effects or item context effects) particularly when constructing more than one form of an assessment using the same items.

Item-order effects are well studied in many domains and with various testing types (Brennan, 1992; Debeer & Janssen, 2013; Hecht, Weirich, Siegle, & Frey, 2015; Leary & Dorans, 1985; Meyers, Miller, & Way, 2009; Sykes & Fitzpatrick, 1992; Yen, 1980). Sources of item-order effects include content priming and preceding item difficulty (Whitely & Dawis, 1976). For content priming, an item-order effect may be present when a preceding item (question A, Form 1, in the example in Figure 1) provides similar or foundational content to the target item (question B in the example), where this same item in another environment may follow an item without such content similarity (question C, Form 2, in the example). An item environment effect could then lead to higher

performance on the target item when the priming item is present.

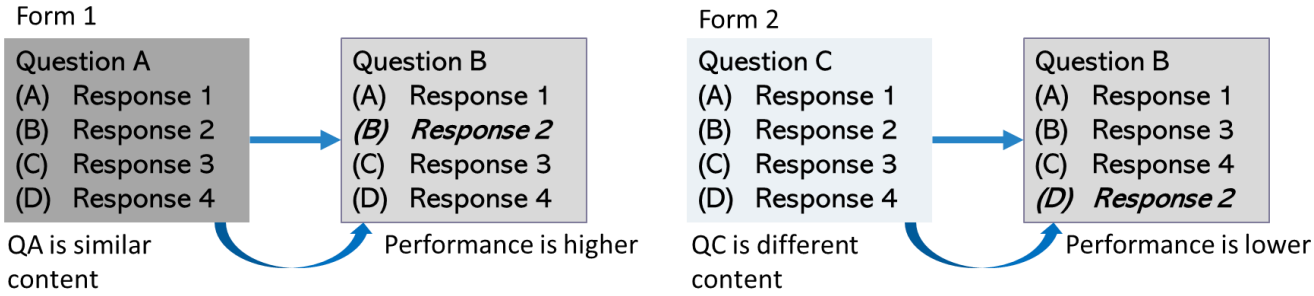
For preceding item difficulty, an item-order effect may be present when a preceding item (question A, Form 1, in the example in Figure 2) is easier compared to the preceding item on a different form (question C, Form 2, in the example). The item environment effect could then lead to higher performance on the target item (again, question B) when the preceding item is easier. This has been seen before (Schroeder et al., 2012) and exact reasons for these effects are not clear; they may be attributed to cognitive factors such as a self-efficacy boost when the previous item is easier and fatigue when previous item is more difficult.

Studies examining these specific item environment effects have found mixed results. For example, some studies have found little to no disadvantages when assessment items are randomized (McLeod, Zhang, & Yu, 2003; Meyers et al., 2009; Schroeder, Murphy, & Holme, 2012). However, it has been found that item environment effects lead to two possible individual item performance differences: priming and preceding item difficulty (Schroeder et al., 2012). In addition to these two categories, item position within a test, *i.e.*, that is whether the item is found earlier versus later in an assessment, has been found (Meyers et al., 2009). Item environment effects have also been studied in the context of item position on trial assessments versus released assessments (Doerner & Calhoun, 2009; Eignor & Stocking, 1986; Huntley and Welch, 1988; Sue, 2009; Whitely & Dawis, 1976). For example, Huntley & Welch found that when items appeared early on trial visions and were then moved to later in the exam on the released version. they were more difficult and visa-versa. The possible influence of item order on overall measurement of student proficiency shown in these studies leads researchers, especially those working large-scale testing programs, to the use of test-equating algorithms (Meyers et al., 2009) to address item environment.

Study Design

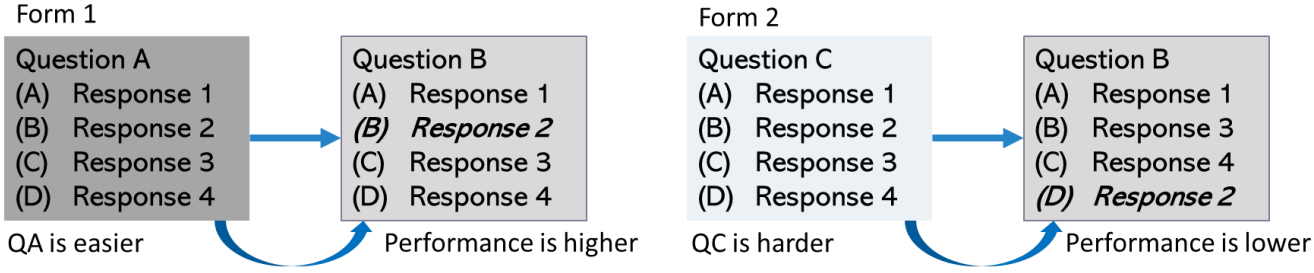
The process by which examinations are developed by the American Chemical Society, Division of Chemical Education, Examinations Institute (ACS Exams) has been previously detailed (Holme, 2003). ACS Exams develops and produces examinations in

Figure 1. Item-order effect of priming example



Note: Item B is the same item on both forms and Response 2 is the correct answer.

Figure 2. Item-order effect of preceding item difficulty example



Note: Item B is the same item on both forms and Response 2 is the correct answer.

all subdisciplines of chemistry for secondary and post-secondary chemistry courses. Examination development typically takes two to three years to complete including a planning, writing, trial testing, and final release stages. The examination development committee typically produces two trial examinations with unique items, developed within the same parameters as the released test (*i.e.*, number of items, timing, content coverage). The committee uses item analysis completed by ACS Exams staff to select the items for the final released examinations. For high-use examinations (such as general chemistry or organic chemistry courses), two forms of the final released test are produced wherein item and response orders are varied.

The issue of item environment effects, methods to detect such effects, and resulting implications are key to the development of quality ACS examinations. Given the static nature of ACS examinations, where the administration of the tests is wholly managed by the institution and instructor, as well as the lifetime of an examinations (released examinations are 'current' for four or more years), it is important to develop robust methods to detect item environment effects as well as address sources of these effects during examination development. ACS Exams has previously used a differential item functioning (DIF) approach based on the Mantel-Haenszel procedure (Holland & Thayer, 1988) when identifying and mitigating item environment issues (Schroeder et al., 2012). The work reported herein expands the use of the DIF approach, particularly examining for differences between three presentations (i.e., trial examination and the two released examination forms) of the item for item environment effects. Our evaluation is based on changes in item parameters modeled by Two-parameter logistic item response theory model (2PL-IRT) (Lord and Novick, 1968). The purpose of this study is to articulate a novel procedure for how 2PL-IRT can be used to evaluate item environment effects on an assessment of cognitive performance and explore possible reasons for the effects. Data from a national assessment on general chemistry will be used to demonstrate the method.

Data

Data used in this study were from the first semester of a yearlong postsecondary general chemistry course. All data were from a research-intensive university in the Midwest of the United States. Item data were collected using four forms of the assessment: two trial examinations (Trial 1, T1, and Trial 2, T2) and two released forms (Released 1, R1, and Released 2, R2); number of responses by form are reported in Table 1. Data reported in Table 1 reflect those examination attempts for which 60 or more items of the 70-item assessment were answered; less that 1% of all responses failed to meet this criterion.

Table 1. Summary of data sets

Form Type	Examination	Number of Respondents
Trial test	T1	529
Trial test	T2	523
Released test	R1	552
Released test	R2	605

Model

Student responses to the items was analyzed using Two-parameter logistic item response theory model (2PL-IRT). In the 2PL-IRT model the probability of a correct response ($X_{is} = 1$) is a function of one student parameter, ability, and two item parameters: discrimination and difficulty.

$$P(X_{is} = 1 | \theta_s, \alpha_i, \beta_i) = \frac{\exp(\alpha_i(\theta_s - \beta_i))}{1 + \exp(\alpha_i(\theta_s - \beta_i))}$$

where θ_s = subject ability,

α_i = item discrimination, and

β_i = item difficulty

This method offers several advantages over the more traditional classical test theory (CTT) method. In CTT, item parameters and the student parameter are sample dependent (Hambleton & Jones, 1993). This means that CTT item parameters will vary by group (e.g., high ability versus low ability). If the data meets the assumptions of IRT, the parameters from this analysis avoid this issue. Results of an 2PL-IRT analysis return item parameters, difficulty and discrimination, and student ability estimates with standard errors. These standard errors allow for the comparison of item performance in different contexts (i.e., assessment forms).

BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) was used to obtain 2PL-IRT item parameter values and student ability estimates. Item parameter

estimates were estimated using BILOG default settings.

Before analysis, all data from each exam administration were checked to ensure that they met the requirements of the 2PL-IRT model: item-model fit, item independence (Yen's Q3 statistic), and parameter invariance. Parameter invariance was checked by splitting each set into two parts and evaluating the correlation between the item parameters, see pages 111-113 in De Ayala, 2009. Data were determined to have satisfactorily met these requirements and were suitable for the reported analyses.

Method

The IRT-based method for identifying item environment effects is comprised of three steps: First, select item on the four assessment forms using items that have consistent item parameters. Second, generate item parameters based on anchor items identified in step 1. Last, probe the items and item environments to identify explanations for the item environment effects using the IRT-based identification method.

Step 1 – Anchor Item Parameters. As data were collected (albeit over several terms) at the same institution with no curricular changes during the data collection timeframe, an approximation to a random groups design (Cook & Eignor, 1991). Trial form data (T1 and T2) were collected during academic terms prior to the released form data (R1 and R2). These conditions, thus, violate the strict assumptions of the random groups design. To account for these violations, an anchoring strategy was used estimate item parameters values on a single scale. A nonequivalent groups anchor design (DeMars, 2014) was used to fulfill this strategy.

Anchor items were considered and selected when their parameters did not display item environment effects. This was done using a DIF analysis in BILOG with a pair wise approach, comparing form A with form B. Specifically, data for items that appear on two forms were combined and used to fit two 2PL-IRT models: i). Assuming no difference in item performance based on form, and ii). Assuming possible difference in item performance based on form. In all comparisons, data has a better fit when possibility of differences were assumed (c.f., (Zimowski et al., 1996) suggesting that DIF was present in all pair wise comparisons of response data.

Lord's *d* (Lord, 1977, 1980), an IRT situated DIF detection method, was used determine item environment effects. The value of *d* for an item is the difference in the estimate of the item's difficulty for the two groups divided by the standard error of the difference. Values of *d* above 1.96 or less than –1.96 are considered significant.

$$d_i = \frac{\beta_2 - \beta_1}{\sqrt{(SE_{\beta_2})^2 + (SE_{\beta_1})^2}}$$

Step 2 - Identification and classification of items displaying form dependence. Item parameters for each examination form were obtained through an IRT calibration using fixed item parameters for the fourteen anchor items via the BILOG NOADJUST option in the Calibration command.[1] As before, Lord's *d* was used to identify form effects, DIF, in the unanchored items. The Example results for a form-independent item parameters are shown in Figure 3. Once form dependent items were identified they were put into categories to investigate possible reasons for the form dependence.

Step 3 – Explanation of the Form Dependencies. Items with form dependent item parameters were analyzed for the degree to which identifiable patterns might explain the observed dependence. These included preceding item difficulty, content priming, and the combination of preceding item difficulty and content priming.

Results

Step 1 – Anchor Item Parameters. Only differences between item difficulty were considered using the BILOG software. Thus, our analysis is limited to detecting uniform DIF, and does not consider non-

[1] Normally, at the end of the analysis, BILOG will rescale the parameters such that the mean ability is zero. Using the NOADJUST in the Calibration command prevents BILOG from rescaling item parameters at the end of the analysis. The NOADJUST option was appropriate (DeMars & Jurich, 2012) due to small (~0.1 logits) in the DIF analysis, scaling parameters for ability distributions, indicating ability distributions for the four forms were similar.

Figure 3: Item parameters displaying no form dependence

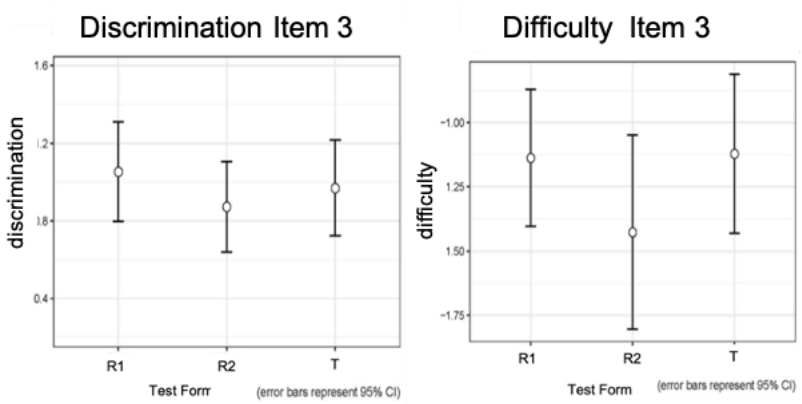


Table 2. Sample values of BILOG output used to test for significant DIF

DIF Study	Item	$\beta_2 - \beta_1$	SE of Diff	d	Sig DIF
R1-T1	1	0.562	0.592	0.950	
R1-R2	1	0.169	0.666	0.254	
R2-T1	1	0.381	0.563	0.677	
R1-T2	9	–1.008	0.164	–6.133	*
R1-R2	9	–0.472	0.131	–3.616	*
R2-T2	9	–0.434	0.136	–3.186	*

uniform DIF (i.e., differences in item discriminations). An example of the three tests of DIF (i.e., trial form versus R1 or R2, and R1 versus R2) are reported for two sample items (Item 1 and Item 9) in Table 2. (Items were mutually exclusive to either T1 or T2; therefore, a comparison between T1 and T2 cannot be made.) Item 1 did not exhibit item environment effects, whereas Item 9 exhibited item environment effects between all three forms that the item appeared on. DIF analyses for all 70 items is reported in Table 3 (items were renumbered for all forms to match form R1); only significant differences between forms are reported for clarity.

Results in Table 3 indicate that there are 34 items suitable for use as anchors (*i.e.*, items that did not display DIF, and thus environment effects). Anchor items were selected using such that they represented ~20% of the total number of items (Woods, 2008), had large item discriminations (Lopez Rivas, Stark, & Chernyshenko, 2008), and item difficulties collectively spanned the range of observed item difficulties (Williams, 1997). Using these criteria, 14 anchor items were selected. Seven of these items were from T1 and seven were from T2 versions of the test. Item parameter values for the 14 anchors are reported in Table 4.

Step 2 - Identification and classification of items displaying form dependence. Of the 56 non-anchor items, 18 items displayed some type of form dependence. These items were organized by differences in difficulties as shown in Table 5 (where the numbering is arbitrary, not preserving the original numbering from any of the four forms) with examples of these shown in Figure 4. These items are classified into three categories using difficulty differences as the predominate characteristic:

- *Category 1* (Items A-I): difficulty dependence based on form where one item was significantly different from the other two. This is shown with the differences in difficulty (consistent with the differences in difficulty (consistent with Figure 4, showing both values) and the highlighted item also shown under difficulty.

Table 3. Results of DIF analysis

Item	R1-R2	R1-T1	R1-T2	R2-T1	R2-T2		Item	R1-R2	R1-T1	R1-T2	R2-T1	R2-T2
1							36					
2							37					
3							38	*				
4							39					
5							40			*		*
6							41		*		*	
7							42				*	
8		*					43	*	*			
9	*		*		*		44					
10							45		*		*	
11			*				46					
12					*		47					*
13							48					
14			*				49					
15	*	*					50	*		*		
16	*			*			51			*		*
17							52	*			*	
18							53		*		*	
19			*		*		54					
20			*		*		55					
21							56			*		*
22		*		*			57				*	
23			*				58					
24							59	*	*		*	
25							60					
26							61			*		*
27				*			62					*
28							63					
29				*			64					
30							65			*		*
31		*		*			66			*		*
32	*		*				67			*		*
33							68			*		*
34							69			*		
35							70					

(Yellow – identified DIF; Gray – comparison not possible; Blue – no DIF identified)

Table 4. Anchor item parameter values

Item	Item Discrimination	Item Difficulty	Trial Form
4	1.388	–1.374	T1
6	1.526	–0.824	T1
18	0.890	–1.312	T2
25	1.331	–1.646	T1
28	1.358	–0.367	T2
35	0.650	0.711	T1
39	0.874	0.0810	T1
44	1.330	–0.672	T1
46	0.902	–0.026	T2
54	0.8165	–0.229	T2
55	0.773	–1.280	T2
60	0.318	0.277	T2
64	1.115	–1.095	T1
70	0.561	–1.379	T2

Table 5. Form dependence items with difficulty and discrimination

Item	Difficulty			Differences in difficulty			Discrimination		
	R1	R2	T	R1-R2	R1-T	R2-T	R1	R2	T
A	–1.537	–2.351	**0.058**		0.650	0.953	0.620	0.388	0.426
B	–0.538	–0.543	**0.945**		0.834	0.845	1.488	1.576	0.510
C	–1.275	–1.295	**0.558**		0.798	0.769	0.800	0.693	0.292
D	–0.464	–0.519	**–1.434**		0.386	0.341	1.003	1.050	0.864
E	-0.162	0.104	**–0.651**		0.113	0.383	0.908	0.884	1.403
F	–1.140	**–1.823**	–1.107	0.209		0.226	2.299	1.428	2.245
G	–0.208	–0.184	**–0.759**		0.137	0.163	0.926	0.893	1.111
H	1.072	**0.251**	1.387	0.015		0.348	0.406	0.804	0.557
I	0.000	–0.020	**–0.450**		0.046	0.044	0.818	0.855	1.092
J	–0.433	–0.788	–1.220		0.324		0.922	1.349	1.203
K	0.231	–0.116	–0.474		0.272		0.832	0.629	0.940
L	–1.672	–0.761	–0.969	0.180			0.707	0.968	0.961
M	0.908	0.987	1.634		0.150		1.443	0.970	1.006
N	–0.718	–0.833	–1.408		0.120		0.986	0.661	1.060
O	–1.123	–0.635	–0.431		0.053		0.807	1.029	0.627
P	–1.241	–1.139	–1.905			0.025	1.079	1.075	1.006
Q	–0.840	–1.076	–1.166		0.020		2.008	2.156	2.320
S	2.795	2.038	3.173				0.512	0.802	0.328

- *Category 2* (Items J-Q): difficulty dependence based on form where the progression of change was not significant, but the highest and lowest difficulty items are significantly different.
- *Category 3* (Item S): discrimination dependence based on form without difficulty dependence based on form.

Category 1 Items

Three examples of an item from Category 1 are shown in Figure 4. In this figure, it can be seen that for Items A and D although the items on the released test (forms R1 or R2) form performed similarly, there was a significant difference. Three examples of these are shown in Figure 5.

Category 2 Items

An additional eight items are classified in category 2 where there was a progression of altering difficulty with only the lowest and highest difficulties as significantly different. Three examples of these are shown in Figure 5.

Category 3 Item

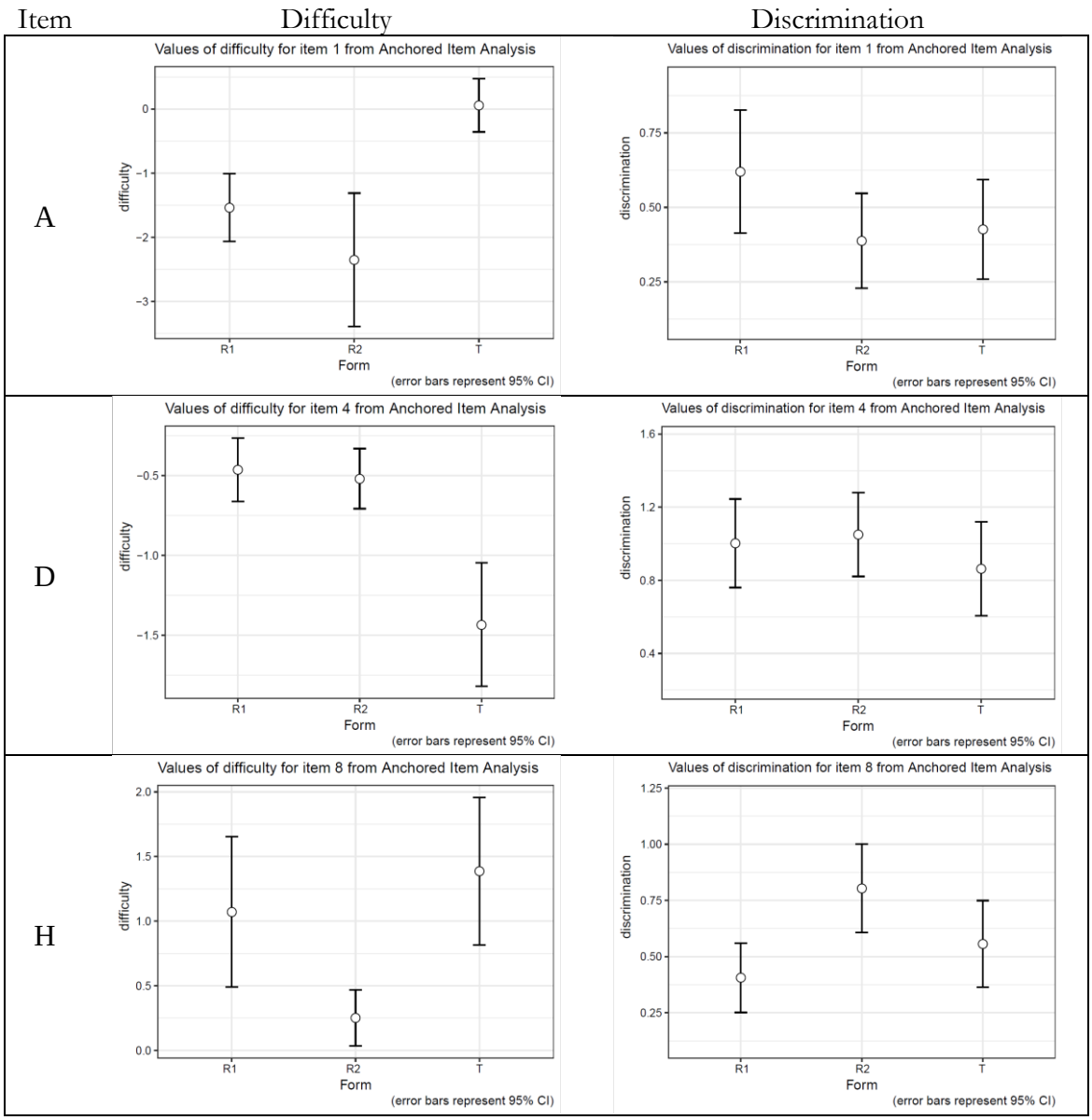
Finally, one item (Item S) significantly differed on item discrimination by form without a significant difference in item difficulty by form; this item is shown in Figure 6.

Step 3 – Explanation of the Form Dependencies

Of the 18 items displaying form dependence, 16 of the items had differences involving item parameters between the trial form and a released form; two items had differences between the two released forms. Items with form dependent item parameters were analyzed for the degree to which identifiable patterns might explain the observed dependence (see Table 6).

Item environment effects fit into four groups. The first three align with effects reported in the literature: I). preceding item difficulty, II). content priming, III). both preceding item difficulty and content priming. A fourth group (IV) did not align with previously found item environment effects. Possible explanations for the observed item environment effects are reported in Table 7 by these four groups.

Figure 4. Category 1 – Singular item difficulty form dependence for three items



The lack of an explanation, for group 4 items, stemming from the two commonly cited reasons for form dependency is initially a concern. However, item S having non-uniform or discrimination only form dependency is unique and unlikely to be explained by these commonly cited reasons. Additionally, there is also evidence to support items placed later on tests have a higher likelihood to exhibit form dependency and although the items varied in their location on the three forms, three out of the four items that were in group 4 for uniform form dependence were on the second half of the form (all occurring at item 41 or higher). Overall, twelve items out of the eighteen found with any form dependence were on the second half of the assessment.

To further investigate the pervasiveness of possible item environment effects, ten items that did not exhibit DIF were selected at random for the same analysis as the form dependent items. These items appeared throughout the examination forms, have a range of difficulties, and are equally distributed across the two trial examination forms. Values for difficulty and discrimination for these non-form dependent items are shown in Table 8.

The possible grouping and explanation for where form dependence could have expected to occur is

Figure 5. Category 2 – Progression of item difficulty form dependence for three items

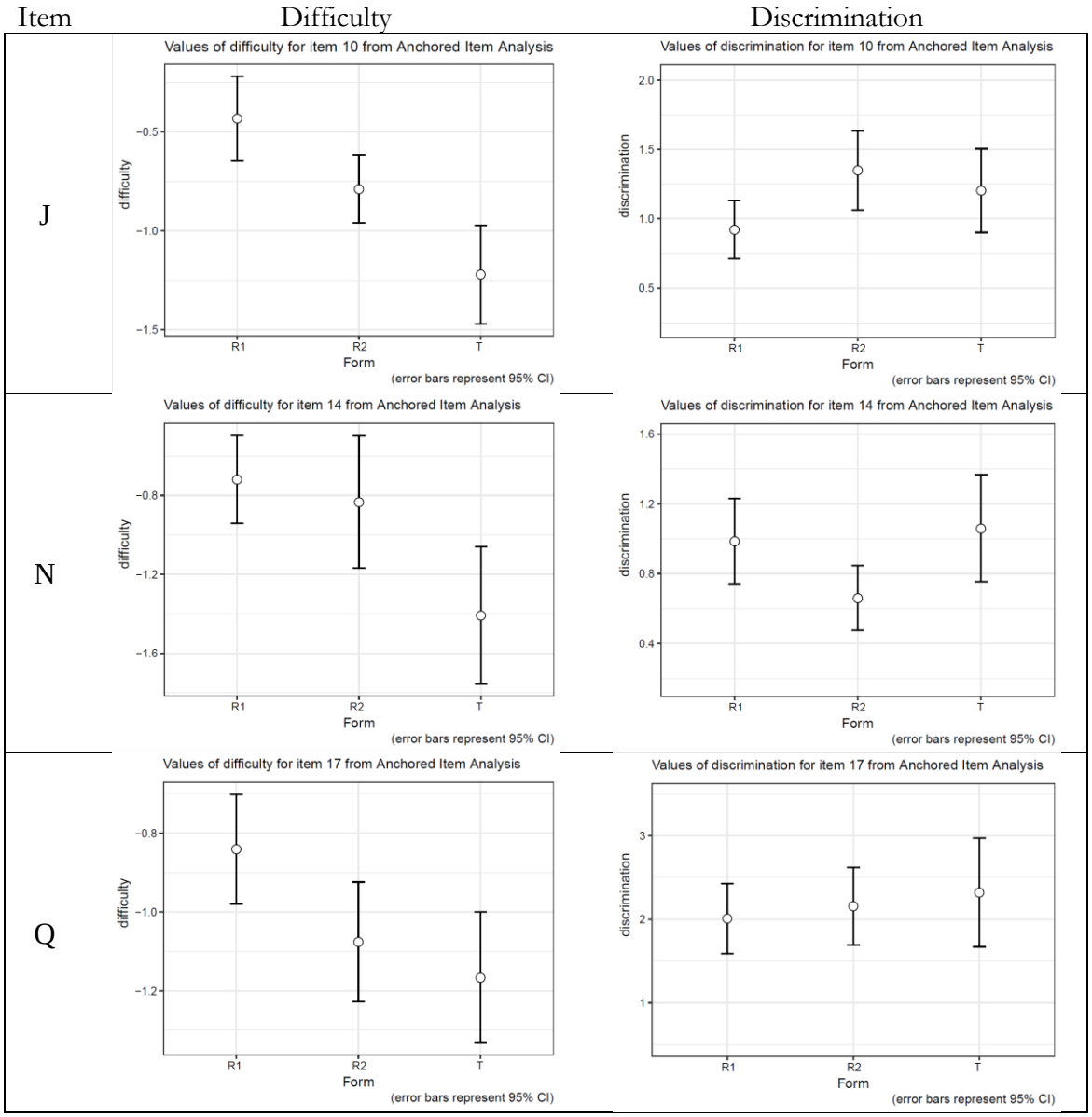


Figure 6. Category 3 – Progression of item discrimination form dependence

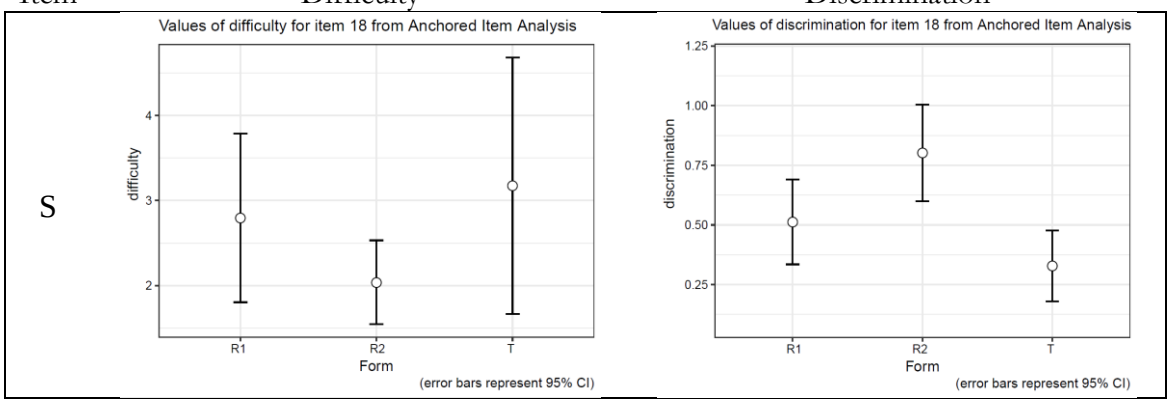


Table 6. Item difficulty and content priming for preceding items that exhibited form dependence with possible reasoning for form dependence

Item	Difficulty			Preceding difficulty			Content priming			Possible reason	
	R1	R2	T	R1	R2	T	R1	R2	T	priming	preceding difficulty
A	–1.537	–2.351	**0.058**	–1.241	0.104	0.558	x	x	x		x
B	–0.538	–0.543	**0.945**	–1.312	–2.738	–0.370					x
C	–1.275	–1.295	**0.558**	–1.537	–1.853	0.058					x
D	–0.464	–0.519	**–1.434**	0.277	0.277	–0.482			x	x	x
E	–0.162	0.104	**–0.651**	–1.214	–1.879	–1.336	x	x	x		
F	–1.140	**–1.823**	–1.107	–0.840	–2.935	–3.602	x				
G	–0.208	–0.184	**–0.759**	–1.095	–1.295	0.015			x	x	
H	1.072	**0.251**	1.387	–0.026	–1.575	1.064		x	x	x	x
I	0.000	–0.020	**–0.450**	0.081	–0.406	–1.154	x	x	x		x
J	–0.433	–0.788	–1.220	–0.599	–1.327	–1.492		x	x	x	x
K	0.231	–0.116	–0.474	–1.280	–0.519	0.277		x	x	x	
L	–1.672	–0.761	–0.969	–1.686	–0.833	–1.779	x	x	x		x
M	0.908	0.987	1.634	0.564	0.120	–0.652					
N	–0.718	–0.833	–1.408	0.000	0.081	0.081	x	x	x		
O	–1.123	–0.635	–0.431	–2.075	–0.672	1.387	x		x	x	x
P	–1.241	–1.139	–1.905	–0.208	1.136	–0.759					x
Q	–0.840	–1.076	–1.166	–0.445	–1.597	–1.696	x		x	x	x
R	2.795	2.038	3.173	–1.328	–1.331	–0.742			x		

shown in Table 9 where Group V is used for no evidence of a reason for form dependency and Group VI for a possible cancelation effect. There were six instances where the selected items fit the "criteria" of one of the categories of form dependence we identified and might be predicted to display some form dependence. However, none of them displayed the form dependence predicted. This suggests that it is only possible to rationalize form dependence items, and not predict its presence.

Conclusions

The purpose of this paper was to articulate a method for using Two-parameter logistic item response theory model to identify item environment effects. Data from a national assessment on general chemistry was used to demonstrate the three step-method. First, identify items that have consistent item parameters and select anchor items from these. Second, generate item parameters for the remaining items using the anchors. Finally, probe the items and item environments to identify explanations for the item environment effects. The method identified 18 of 56 non-anchor items that displayed item environment effects. Thirteen of the 18 identified items were explained by reasons previously reported in the research literature: preceding item difficulty, content priming, and both preceding item difficulty and content priming. Five remaining items were unable to be sufficiently explained.

Ten non-anchor items for which item environment effects were not found were randomly selected and evaluated for predicted item environment effects and associated reasonings. This process is akin to looking for counterexamples.

For example, if an item immediately following a hard item on form A had a lower difficulty on form B where it immediately followed an easier item (preceding item difficulty dependence), item environment effects would be predicted. We found six of the ten items "should" have displayed one of the three types of form dependence. The inability predict item environment effects is disappointing. If it were

Table 7: Groupings and possible explanations of form dependence

Item	Group	Possible explanation
A	I	Preceding item on T was harder, leading to lower performance on target item. *
B	I	Preceding item on T was easier, leading to higher performance on target item.
C	I	Preceding item on T was harder, leading to lower performance on target item.
I	I	Preceding item on T was easier, leading to higher performance on target item. *
L	I	Preceding item on R1 was easier, leading to higher performance on target item. *
P	I	Preceding item on T was easier, leading to higher performance on target item.
G	II	Priming present for T where performance was highest on target item.
K	II	Priming present for T and R2 where performance was highest on target item for T.
D	III	Preceding item on T was easier, leading to higher performance on target item; content priming present for T only.
H	III	Preceding item on R2 was easier, leading to higher performance on target item; content priming present for R2 and T.
J	III	Preceding item on T was easier, leading to higher performance on target item; content priming present for T and R2.
O	III	Preceding item on R1 was easier, leading to higher performance on target item; content priming present for R1 and T.
Q	III	Preceding item on T was easier, leading to higher performance on target item; content priming present for R1 and T.
E	IV	R2 had the easiest preceding item, so should have also had the highest performance on target item, however T was highest. *
F	IV	T was easier than R2 for preceding item with content priming for R1. Either R1 or T should have performed higher than R2.
M	IV	T had the easiest preceding item, so should have also had the highest performance on target item, however T was lowest.
N	IV	No clear pattern discernable.
S	IV	Item was very difficult; unable to discern source of non-uniform form dependence.

* content priming present for all three items, so no effect.

Table 8: Item difficulty and content priming for preceding items that did not exhibit form dependence

Position in test	Trial form	Difficulty			Preceding difficulty			Content priming			Priming	Preceding difficulty
		R1	R2	T	R1	R2	T	R1	R2	T		
early	T1	–2.179	–2.368	–2.291	–1.374	–2.087	–1.589					
early	T2	–1.105	–1.196	–1.563	–2.701	–0.788	–1.222					x
early	T2	–0.445	–0.676	–0.746	–3.039	–1.196	–1.563		x	x	x	x
middle	T1	0.463	0.431	0.675	2.795	–0.543	3.173	x		x	x	x
middle	T1	–1.278	–1.515	–1.025	–0.370	1.158	–2.246	x			x	x
middle	T2	–0.557	–0.785	–0.877	–1.089	–0.985	0.236					x
late	T1	0.237	0.120	0.545	–0.672	–0.026	1.634					x
late	T2	1.310	1.270	1.064	–1.123	0.987	–0.125	x			x	x
late	T1	1.027	1.136	0.794	0.114	–0.184	1.789		x		x	
late	T2	–1.539	–1.853	–2.318	–1.275	–2.351	–1.629		x		x	x

Table 9: Groupings and possible explanations of lack of form dependence

Position in test	Trial form	Group	Possible explanation
early	T1	V	Fairly equivalent preceding difficulties; no priming
early	T2	I	Given preceding difficulty on R1, would have expected higher performance on target item.
early	T2	VI	Possible cancelation effect with easy preceding item for R1; priming for R2 and T.
middle	T1	VI	Possible cancelation effect with easy preceding item for R2; priming for R1 and T.
middle	T1	VI	Possible cancelation effect with easy preceding item for T; priming for R1.
middle	T2	I	Given preceding difficulty on R1, would have expected higher performance on target item.
late	T1	I	Given preceding difficulty on R1, would have expected higher performance on target item.
late	T2	III	Given preceding difficulty on R1 and priming, would have expected higher performance on target item.
late	T1	III	Given preceding difficulty on R2 and priming, would have expected higher performance on target item.
late	T2	III	Given preceding difficulty on R2 and priming, would have expected higher performance on target item.

possible to predict instance when items would "misbehave", this could be modeled and accounted for when constructing an assessment. This would benefit assessment developers as various examinations forms are constructed from a pool of assessment items.

The 2PL-IRT method reported herein adds to and compliments other methods used to identify unstable assessment items. We note that the robust z statistic, extended to 3 parameter models and 2 parameter partial credit models, can be used to investigate the variations in linking parameters that could identify unstable items (Huynh & Meyer, 2010). Our method, though, differs from this in that we do not examine linking parameters; instead, in using anchor items, we examine changes in the item parameters directly. The Mantel-Haenszel statistic has been used in an approach similar to ours to identify instability in items common to multiple forms of an assessment (Michaelides, 2008). While this approach is similar to ours, it does not take advantage of the benefits offered by using an item response approach, i.e., the ability to use item parameter standard errors to identify misbehaving items. Future work will compare our new method with an analysis using the Mantel-Haenszel approach.

Limitations

Some important limitations must be mentioned. This work was done using existing items that appeared in test forms that were given at different times and to different students. This resulted in using an item-anchoring procedure that has features of both an equivalent groups design and a nonequivalent anchor groups design. However, the pairwise DIF analysis gave ability scaling factors (necessary to put the ability estimates on the same scale) of less than 0.1 logits, which indicates that the underlying ability distributions of the four groups of students responding, two released versions and two trial versions, were similar. This indicated that our data could be analyzed using the techniques of the equivalent group design procedure. All attempts were made to be as conservative as possible in the anchoring process, but the interpretation of the results of this study should consider these limitations.

References

Brennan, R. L. (1992). Context of context effects. *Applied Measurement in Education*, *5*(3), 225–264.

Cook, L. L., & Eignor, D. R. (1991). IRT Equating Methods. *Educational Measurement: Issues and Practice*, *10*(3), 37–45.

De Ayala, R. J. (2009). Theory and Practice of Item Response Theory. New York: The Guilford Press.

De Ayala, R. J. (2010). Item Response Theory. In G. R. Hancock & R. O. Mueller (Eds.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (pp. 155–172). New York: Routledge.

Debeer, D., & Janssen, R. (2013). Modeling Item-Position Effects Within an IRT Framework.

Journal of Educational Measurement, *50*(2), 164–185. http://doi.org/10.1111/jedm.12009

DeMars, C. E. (2014). A comparison of confirmatory factor analysis and multidimensional Rasch models to investigate the dimensionality of test-taking motivation. *Journal of Applied Measurement*, *14*(2), 179–196.

DeMars, C. E., & Jurich, D. P. (2012). Software Note. *Applied Psychological Measurement*, *36*(3), 232–236. http://doi.org/10.1177/0146621612438726

Doerner, W. W., & Calhoun, J. P. (2009, April 2). The Impact of the Order of Test Questions in Introductory Economics.

Eignor, D. R., & Stocking, M. L. (1986). *An Investigation of Possible Causes for the Inadequacy of IRT Pre-Equating* (No. RR-86-14). *ETS Research Report* (pp. 1–50).

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47.

Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Modeling Booklet Effects for Nonequivalent Group Designs in Large-Scale Assessment. *Educational and Psychological Measurement*, *75*(4), 568–584. http://doi.org/10.1177/0013164414554219

Holland, P. W., & Thayer, D. T. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In H. Wainer (Ed.), *Test Validity* (pp. 129–145).

Holme, T. A. (2003). Assessment and Quality Control in Chemistry Education. *Journal of Chemical Education*, *80*(6), 594–596. http://doi.org/10.1021/ed080p594

Huntley, R. M., and Welch, C. J. (1988). Numerical Answer Options: Logical or Random Order? *Annual Meeting of the American Educational Research Association*.

Huynh, H., & Meyer, P. (2010). Use of Robust z in Detecting Unstable Items in Item Response Theory Models. *Practical Assessment, Research & Evaluation*, *15*, 1–9. http://doi.org/10.7275/ycx6-e864

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, *55*(3), 387–413.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2008). The Effects of Referent Item Parameters on Differential Item Functioning Detection Using the Free Baseline Likelihood Ratio Test. *Applied Psychological Measurement*, *33*(4), 251–265. http://doi.org/10.1177/0146621608321760

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), Basic problems in cross-cultural psychology (pp. 19-29). Amsterdam: Swets and Zeitlinger.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley.

McLeod, I., Zhang, Y., & Yu, H. (2003). Multiple-choice randomization. *Journal of Statistics Education*, *11*(1).

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item Position and Item Difficulty Change in an IRT-Based Common Item Equating Design. *Applied Measurement in Education*, *22*(1), 38–60. http://doi.org/10.1080/08957340802558342

Michaelides, M. P. (2008). An Illustration of a Mantel-Haenszel Procedure to Flag Misbehaving Common Items in Test Equating. *Practical Assessment, Research & Evaluation*, *13*, 1–17. http://doi.org/10.7275/n04d-8767

Rogers, H. J., & Swaminathan, H. (1993). A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement*, *17*(2), 105–116.

Schroeder, J., Murphy, K. L., & Holme, T. A. (2012). Investigating Factors That Influence Item Performance on ACS Exams. *Journal of Chemical Education*, *89*(3), 346–350. http://doi.org/10.1021/ed101175f

Sue, D. L. (2009). The Effect of Scrambling Test Questions on Student Performance in a Small

Class Setting. *Journal for Economic Educators*, *9*(1), 32–41.

Sykes, R. C., & Fitzpatrick, A. R. (1992). The Stability of IRT b Values. *Journal of Educational Measurement*, *29*(3), 201–211.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of Differential Item Functioning Using the Parameters of Item Response Models. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67–113). Hillsdale, NJ: Routledge.

Whitely, S. E., & Dawis, R. V. (1976). The Influence of Test Context on Item Difficulty. *Educational and Psychological Measurement*, *36*, 329–337.

Williams, V. S. L. (1997). The "Unbiased" Anchor: Bridging the Gap Between DIF and Item Bias. *Applied Measurement in Education*, *10*(3), 253–267. http://doi.org/10.1207/s15324818ame1003_4

Woods, C. M. (2008). Empirical Selection of Anchors for Tests of Differential Item Functioning. *Applied Psychological Measurement*, *33*(1), 42–57. http://doi.org/10.1177/0146621607314044

Yen, W. M. (1980). The Extent, Causes, and Importance of Context Effects on Item Parameters for Two Latent Trait Models. *Journal of Educational Measurement*, *17*(4), 297–311.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multigroup IRT Analysis and test maintenance for binary items. Chicago: Scientific Software.

Citation:
Pentecost, T. C., Raker, J. R., & Murphy, K. L. (2023). Application of two-parameter item response theory for determining form-dependent items on exams using different item orders. *Practical Assessment, Research, & Evaluation*, 28(9). Available online: https://scholarworks.umass.edu/pare/vol28/iss1/9/

Corresponding Author:
Kristen L. Murphy
University of Wisconsin - Milwaukee
Email: kmurphy@uwm.edu