Using Item Difficulty and Item Position to Measure Test Fatigue

Jeff Davis (AIR) • Abdullah Ferdous (University of Nebraska, Lincoln)

American
Institutes
for Research ®

1000 Thomas Jefferson Street, NW | Washington, D.C. 20007-3835 | 202.403.5000 | www.air.org

Table of Contents

ABSTRACT

With the increase in standardized testing to address accountability issues, educators are becoming more concerned with whether there is a student test fatigue effect on state assessments. If students are not performing as well on items appearing later on tests, efforts to accurately assess performance standards may be compromised. In addition, longer tests may have a differential impact on disadvantaged students. This study compared item difficulties (p-values and b-parameters) from administering the same items in field testing and live testing. Negative changes in item performance based on test position may indicate student fatigue. Contrary to some previous evidence, results showed declines in student performance as items dropped to lower positions on tests.

THEORETICAL FRAMEWORK

Many educators are concerned with whether there are test fatigue effects on their state assessments, both in terms of the number of tests and the length of tests. The main issue with this study is whether perceived concerns with test length among educations are supported by empirical evidence. If the effects are real, test fatigue would negatively impact efforts in measuring achievement. If not, backlashes against testing programs and accountability measures would be unjustified.

Previous research has shown mixed results on the effects of item positioning on examinee performance. Some studies have reported a decline in performance as items appear later on tests (PISA, 2000, Wise, Chia, & Park, 1989) whereas others reported no differences in item difficulty and test position (Rubin & Mott, 1984; Klein & Bolus, 1983; Zwick, 1991). In a study with special education students, performance on tests was actually lower when efforts were made to reduce fatigue by giving students parts of tests in multiple sessions rather than in a single sitting (Walz, et. al., 2000). No studies were found in the literature that disaggregating students by group membership (e.g., racial/ethnic or language proficiency).

METHODS

This research study investigates whether there are any differences in examinee performance on the same items administered to randomly equivalent groups during field testing and later during live testing, with different test positions in the two administrations. The assumption is that stability in item difficulty would suggest no effect of item location on examinee performance; i.e., an indication of the absence of test fatigue.

Two different analyses were conducted to investigate the effect of item position on student performance. Analysis 1 evaluated the stability of item difficulty estimates using analysis of variance. Analysis 2 estimated correlation coefficients for the item difficulty estimates. The steps involved in the two methods of analysis are briefly described below. A diagram is provided in Figure 1.

ANALYSIS 1

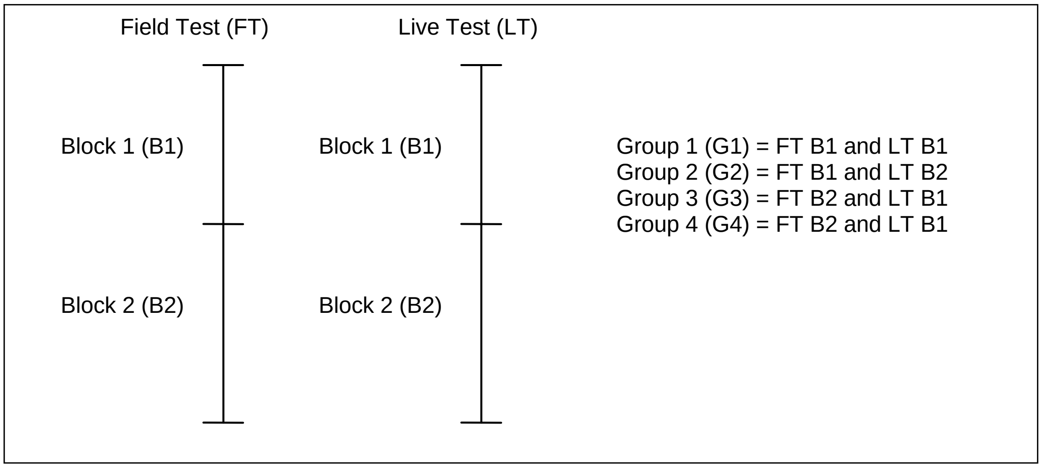
1. Items were classified into two blocks of equal length on both the field tests and live tests with the same items. The first half of the items on the test constituted Block 1 and items in the second half of tests formed Block 2.

2. Four groups of items were identified from these item blocks: Group 1 consisted of items that appeared in Block 1 both in field and live test; Group 2 consisted of items that were in Block 1 of the field test but appeared in Block 2 of the live test; Group 3 contained items that were in Block 2 of the field test and in Block 1 of the live test; and Group 4 consisted of items that were in Block 2 in both the field and live tests.

3. Average item difficulty (p-values and b-parameters) were estimated for each of item groups.

4. The average item difficulty and b-parameters were compared among the groups (LT- FT) using one-way ANOVA followed by a post-hoc analysis (if necessary).

Figure 1. Item Positions and Groups


Field Test (FT)
Live Test (LT)
Block 1 (B1)
Block 1 (B1)
Block 2 (B2)
Block 2 (B2)
Group 1 (G1) = FT B1 and LT B1
Group 2 (G2) = FT B1 and LT B2
Group 3 (G3) = FT B2 and LT B1
Group 4 (G4) = FT B2 and LT B1

ANALYSIS 2

1. Item difficulty (p-values) and b-parameter differences between the field test and the live test were estimated (LT – FT).

2. Position differences for the items on the field and live tests were calculated (FT – LT).

3. Pearson correlation coefficients were estimated between: a) item p-values on the field test and its position difference (between the field and live tests), and b) item b-parameters of the field test items and its position difference (between the field and live tests).

4. Statistical tests were conducted on each of the coefficients.

DATA SOURCES

Results from the grades 3 and 5 tests on a standardized state-testing program were used. Approximately 30,000 students took each of the tests. The tests included core (live) items and embedded field-test (FT) items. For the field-testing, there were five versions of each test, with about 6,000 students taking each set of field test items. Versions were spiraled within classrooms. The items that were field tested in 2001 through 2003 and then appeared as live items on the Spring 2004 operational forms. Tables 1 and 2 provide the structure of the tests by grade level and subject, with the number of live and FT items per form.

Table 1. Test Structure (Field Tests)

Grade	Subject	Number of Items per Form (Live Test and Field Test)		
		Spring 2001	Spring 2002	Spring 2003
3	Reading	--	55	65
	Math	--	60	70
5	Reading	--	53	67
	Math	--	54	74

Table2 Test Structure (Live Tests)

Grade	Subject	Number of Items per Form (Spring 2004)		
		Live	FT	Total
3	Reading	40	25	65
	Math	45	25	70
5	Reading	45	22	67
	Math	49	25	74

RESULTS

Classical statistics and item response theory parameters were calculated for the items. For the IRT analysis, the items were calibrated using the 3-parameter logistic model and placed on the same scale.

GRADE 3 READING

The analysis of variance conducted on the p-value and b-parameter differences (Analysis 1) showed no statistically significant differences (Table 3). No post-hoc tests were conducted (Table 3). In addition, none of the Pearson correlation coefficients were statistically significant (Analysis 2). The conclusion for Grade 3 Reading was that there were no fatigue effects based on the lack of a relationship between item p-value and b-parameter differences and item positions on the tests.

Table 3. Grade 3 Reading

Analysis	Group	Number of Items	Mean P-Value			Mean B-Parameter		
			FT	LT	Diff	FT	LT	Diff
1	G1 (B1 to B1)	15	0.73	0.73	0.00	-0.42	-0.73	-0.32
	G2 (B1 to B2)	8	0.56	0.55	-0.01	0.26	0.16	-0.10
	G3 (B2 to B1)	8	0.63	0.63	0.00	-0.03	-0.20	-0.17
	G4 (B2 to B2)	9	0.61	0.60	-0.01	0.01	-0.07	-0.09
	Overall	40	0.65	0.64	0.00	-0.11	-0.30	-0.19
2	Correlations	40			0.06			-0.10

GRADE 3 MATH

For Grade 3 Math, the analysis of variance (Analysis 1) also showed no significant differences between the groups (Table 4). The second analysis, however, did result in statistically significant Pearson correlation coefficients for the relationship between the size of the position difference and the p-value and b-parameter differences (Analysis 2). Note that the correlation with the p-values was positive (with items that moved forward on the live tests having higher p-values on those tests) and the correlation with the b-parameters was negative (since, unlike p-values, lower b-parameters are associated with easier items).

Table 4. Grade 3 Math

Analysis	Group	Number of Items	Mean P-Value			Mean B-Parameter		
			FT	LT	Diff	FT	LT	Diff
1	G1 (B1 to B1)	14	0.72	0.70	-0.01	-0.48	-0.67	-0.19
	G2 (B1 to B2)	7	0.72	0.70	-0.02	-0.62	-0.78	-0.16
	G3 (B2 to B1)	13	0.75	0.76	0.01	-0.65	-0.78	-0.13
	G4 (B2 to B2)	11	0.69	0.68	-0.01	-0.48	-0.69	-0.21
	Overall	45	0.72	0.71	0.00	-0.55	-0.77	-0.22
2	Correlations	45			0.34*			-0.32*

* represents a statistically significant value at p < .05

GRADE 5 READING

The mean p-value and b-parameter differences resulted in statistical significance, with post hoc differences involving G2 (with G1, G3, and G4). Note that G2 involved items moving from higher positions on the field tests to lower positions on the live test, resulting in lower p-values and higher b-parameters. Both of the Pearson correlations were significant (Table 5).

Table 5. Grade 5 Reading

Analysis	Group	Number of Items	Mean P-Value			Mean B-Parameter		
			FT	LT	Diff	FT	LT	Diff
1	G1 (B1 to B1)	13	0.67	0.67	0.00	-0.32	-0.47	-0.15
	G2 (B1 to B2)	14	0.67	0.58	-0.09*	-0.33	-0.06	0.27*
	G3 (B2 to B1)	11	0.63	0.60	-0.03	-0.16	-0.20	-0.04
	G4 (B2 to B2)	7	0.64	0.62	-0.02	-0.26	-0.23	0.03
	Overall	45	0.66	0.62	-0.04	-0.27	-0.24	0.03
2	Correlations	45			0.63*			-0.65*

* represents a statistically significant value at p <.05

GRADE 5 MATH

The Grade 5 Math analysis resulted in no significant ANOVAs. However, both of the correlations were significant, with more difficulty as the items moved from the beginning positions on the field tests to the end positions of the live test and vice versa (Table 6).

Table 6. Grade 5 Math

Analysis	Group	Number of Items	Mean P-Value			Mean B-Parameter		
			FT	LT	Diff	FT	LT	Diff
1	G1 (B1 to B1)	14	0.69	0.67	-0.02	-0.55	-0.52	0.02
	G2 (B1 to B2)	12	0.60	0.59	-0.01	0.00	-0.05	-0.06
	G3 (B2 to B1)	12	0.60	0.61	0.02	-0.07	-0.30	-0.23
	G4 (B2 to B2)	11	0.65	0.64	-0.01	-0.27	-0.33	-0.06
	Overall	49	0.64	0.63	-0.01	-0.23	-0.31	-0.08
2	Correlations	49			0.43*			-0.45*

* represents a statistically significant value at p <.05

CONCLUSIONS

The purpose of this research study was to investigate the effects of test fatigue. Data from student performance in two different grades levels on a state assessment were used. Proxies for test fatigue were calculated comparing differences in item difficulty (p-values and b-parameter estimates) between the field and the live test. Any statistically significant differences would suggest that there was some influence of test fatigue on student scores.

Two analyses were carried out for each of the grade levels and subject areas. Analysis 1 was used to find overall mean differences in p-values and b-values between the field and the live test item blocks. Analysis 2 investigated the difference in p-value and b-parameter between the field and the live test through Pearson correlations.

The results from Analysis 1 (comparing changes in item difficulties by movement of items between blocks on the tests) showed only that the relationship between item position and item difficulty on Grade 5 Reading was statistically significant (at $p < 0.05$). However, from Analysis 2, the correlation analyses for all tests except for Grade 3 Reading showed statistically significant results, indicating that there was a relationship between item repositioning on the tests and item difficulty.

The differences in conclusions might have occurred due to the limitations of Analysis 1, which converted a continuous variable (item test position) into a binary variable (Block 1 or 2). In most situations, there is a loss of information when a continuous variable is converted into a categorical variable, resulting in statistical tests that may not find actual differences. The recommendation from the results of this study is to place more importance on the correlational analysis, though the binary analysis and the ANOVAs may be easier to understand.

EDUCATIONAL IMPORTANCE

The results from this analysis generally support the concerns of many educators who are apprehensive about whether there are fatigue effects on their state assessments due to the length of the tests. When designing tests, the key is to find a balance between having enough items to reliably measure a state’s standards (as well as repopulate the item bank) and at the same time not overburden students who may become overly fatigued towards the end of tests. The factors of measurement validity and reliability are crucial, but this research should provide a contribution to the debate on whether the length of tests has an effect on student performance.

A recommended next step in this type of analysis study would be to provide more depth into the test length issue by conducting disaggregated analyses. These could involve students by ability level, gender, and racial/ethnic group. Further research may also involve the effects of issues such as language proficiency and the need for accommodations, in addition to developing other methods for measuring item performance by test position.

REFERENCES

Klein, S. P., & Bolus, R. (1983). *The effect of item sequence on bar examination scores*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec.

PISA. (2000). *PISA 2000 Technical Report – Magnitude of Booklet Effects*. Paris, France: OECD Programme for International Student Assessment (PISA).

Rubin, L. S., & Mott, D. E. (1984). *The effect of the position of an item within a test on the item difficulty value.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Walz, L., Albus, D., Thompson, S., & Thurlow, M. (2000). *Effect of a multiple day test accommodation on the performance of special education students*. Minneapolis, MN: National Center on Educational Outcomes, Report 34.

Wise, L. L., Chia, W. J., & Park, R. (1989). *Item position effects for test of word knowledge and arithmetic reasoning*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practices, 10*(3), 10-16.