Item Difficulty Anchoring Procedure for the ASPECt Project

Participants and Data Collection

Items were administered to students during three separate studies in the spring of 2015, winter of 2015-2016, and fall of 2016. Students in various grade bands particated in testing, including grades 4th-5th, 6th-8th, 9th-12th, and Univeristy/College students. Students were adminstered items as either a paper-based test (PBT) or computer-based tests (CBT). Over the course of the project several different computer-based test modes were used. Each computer mode had different functionalities which were tested for whether they resulted in comparable measures to paper-and-pencil testing.

Table 1: *Summary of Student Deomographics for the Full Data Set*

		Spring 2015	**Winter 2016**	**Fall 2016**
Grade Band	4th-5th	2989	1370	1667
	6th-8th	10472	5776	4601
	9th-12th	7493	5284	4511
	University/College	0	574	0
Gender	Female	10466	6666	5564
	Male	10137	5678	4695
Primary Language	English	18242	11360	9533
	Not English	2287	1018	741
Test Mode	Paper Version	6643	4267	3190
	Computer Version 1	14418	0	0
	Computer Version 2	0	8740	0
	Computer Version 3	0	0	2535
	Computer Version 4	0	0	2527
	Computer Version 5	0	0	2527

Creation of the Anchoring Data Set

Item measures were anchored using an anchoring data set which consisted of a subset of the pilot test data. The anchoring data set was created by removing students who (1) may have performed differently due to the test mode effects and (2) only answered a small number of test questions. A comparability study of the PBT and CBT modes indicated some CBT modes could result in lower student performance. Based on this analysis, only students who took the PBT, CBT-1, and CBT-3 test modes were included in the anchoring data set. In addition to removing students due to mode effects, we removed students who answered less than 6 out of 35 items (~17%). A total of 14,036 students were removed from the field test data, resulting in the anchoring data set consisting of 30,811 students. Table 2 summarizes the demographic information for the anchoring data set.

Table 2: *Summary of Student Demographic for the Anchoring Data Set*

		Spring 2015	**Winter 2016**	**Fall 2016**	**Total**
Grade Band	4th-5th	2967	470	848	14%
	6th-8th	10390	1651	2425	47%
	9th-12th	7414	1895	2408	38%
	University/College	0	244	0	1%
Gender	Female	10375	2239	2971	52%
	Male	10052	1952	2494	48%
Primary Language	English	18090	3639	5022	91%
	Not English	2263	514	427	9%
Test Mode	Paper Version	6628	4260	3172	46%
	Computer Version 1	14242	0	0	46%
	Computer Version 2	0	0	0	
	Computer Version 3	0	0	2509	8%
	Computer Version 4	0	0	0	
	Computer Version 5	0	0	0	

Rasch Analysis of Anchoring Data

Rasch analysis was used to estimate item difficulties and student measures from the anchoring data set. In the Rasch model, the probability of the n^{th} student answering the i^{th} item correct, P_{ni}, is related to difference between the student's measure, θ_n, and the item's difficulty, D_i through a logistic function.

$$P_{in} = \frac{e^{(\theta_n - D_i)}}{1 + e^{(\theta_n - D_i)}}$$

Student measures and item difficulties for the Rasch model were estimated using the software package WINSTEPS (Linacre, 2016). Table 3 shows the fit statistics of Rasch model to the anchoring data set.

Table 3: *Summary of Rasch Fit Statistics*

	Item			Student		
	Min	Max	Median	Min	Max	Median
Standard error	0.02	0.10	0.06	0.35	1.93	0.4
Infit mean-square	0.85	1.28	0.99	0.46	2.17	0.99
Outfit mean-square	0.69	1.65	0.99	0.23	5.33	0.98
Point-measure correlation	-0.04	0.54	0.35	-0.94	0.94	0.32
Separation index (Reliability)	13.23 (0.99)			1.51 (0.70)		

Two items were found to have outfit mean-square values greater than 1.4, indicating unexpected responses to these items, and two items had point-measure correlations less than zero, indicating their score responses may not correlate with student knowledge. The fit statistics for these items suggested that some students were unexpectaly getting items correct, possibily due to guessing.

One technique addressing student guessing is to assume guesing is a function of the difficulty of the item and the proficiency of the student. If a low profiecency student gets a high difficulty items correct, it could be a sign that the student correctly guessed. In the Rasch model, this information is captured in students z-residuals for an items, where a high z-residual indicates the students was unlikely to answer the item correctly but they actually did. To decrease the influence of guessing on our item measures we used an approach outlined by Andrich *et al.* (2012), in which a tailored data set is created by replacing student responses with large z-residual values with missing data We replaced all student response's with z-residuals greater than 4 with missing data resulting in a total of 648 responses being replaced with missing data. Table 4 shows the fit statistics after these responses were replaced. A table of the properties for all of the items can be found in Appendix A.

Table 4: *Summary of Rasch Fit Statistics after replacing responses with large z-residuals*

	Item			Student		
	Min	Max	Median	Min	Max	Median
Standard error	0.02	0.10	0.06	0.35	1.93	0.4
Infit mean-square	0.85	1.28	0.99	0.44	2.2	0.99
Outfit mean-square	0.68	1.40	0.99	0.20	3.34	0.98
Point-measure correlation	0.00	0.53	0.35	-0.94	0.94	0.32
Separation index (Reliability)	13.49 (0.99)			1.56 (0.71)		

For the tailored data set all items had Infit and Outfit statistics in an acceptable range (0.6 < Infit, Outfit < 1.4), had positive point-measure correlations, and high separation index. We also tested items for unidimensionality by performing a Principle Component Analysis (PCA) on the items' standardized residuals using WINSTEPS. More than 20% of the variance in the data was explained by the model and eigenvalue of the first contrast was less than two (1.85) providing evidence that the unidimensionality assumption holds (Chaou and Wang, 2010; Linacre, 2016) Overall, the fit statistics indicated items were unidimensional, precisely located on the latent variable, and item hierarchy is accurate.

Differential Item Functionality

To validate that the items were fair we looked for measurement bias for subgroups of students by conducing differential item functioning (DIF) analysis. DIF analysis was conducted to analyze the fairness of items for students of different genders and for students who indicated English was not their primary language. Items were flagged if they were found to have slight to moderate or moderate to large DIF based on their DIF contrast and Mantel-Haenszel statistics (Zwick, 2012). Three items were found to have slight to moderate DIF when comparing gender and nineteen items were found to have slight to moderate DIF when comparing students whose primary language was English to students who indicated English was not their primary language. No items were found to have moderate to large DIF.

To further examine functionality of items with slight to moderate DIF we conducted a second stage of DIF analysis and reviewed text and context of each item. A second Rasch analysis was conducted based on the approach outlined by Zenisky and Hambleton (2003). In this second analysis, items which were previously flagged for DIF were not included in estimating item or person measures. Out of the twenty two items flagged for having slight to moderate DIF initially, nineteen were flagged in the second analysis (three for gender, sixteen for primary language). These nineteen items were reviewed to see whether the context or language in the item could cause the item to be more or less difficult to a subgroup of students. We did not find any item characterisitics that would explain why three items were flagged for slight to moderate gender DIF or primary language DIF. Because of this we did not remove these items from the item bank.

References

Andrich, D., Marais, I., & Humphry, S. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics, 37*(3), 417-442.

Chaou, Y. T, & Wang, W. C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement, 70*(4), 717-731.

Linacre, J. M. (2016). Winsteps® Rasch measurement computer program. Beaverton, Oregon. Retrieved from Winsteps.com

Zenisky and Hambleton (2003). Detection of differential item functioning in large-scale state assessments: a study evaluating a two-stage approach. *Educational and Psychological Measurement, 63*(1), 51-64.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement*. (Tech. Rep., Research Report No. RR-12-08). Princeton, NJ: Educational Testing Service.