The Marginal Utility of Interim Assessments for Prediction

Working Paper

July, 2017

Nathan Dadey

Center for Assessment

Abstract

One key selling point for interim, or benchmark, assessments is that their scores are generally highly predictive of scores on large-scale summative assessments. This paper investigates whether prediction is a useful criterion in justifying the use of interim assessments. To do so, we use ordinary least squares and logistic regression to predict scale scores and proficiency classifications on a sixth grade, large-scale summative mathematics assessment using scores from (a) three sixth grade mathematics interim assessments, which were administered quarterly, and (b) a large-scale summative assessment administered in the prior school year (i.e., fifth grade). We show that scores from the fifth grade summative assessment predict performance on the sixth grade assessment similarly, or very slightly worse, than any interim assessment.

The Marginal Utility of Interim Assessments for Prediction

Introduction and Purpose

One key selling point for interim, or benchmark, assessments is that scores on these assessments are highly predictive of scores on large-scale summative assessments, often administered at the end of the school year. The purpose of this paper is to investigate whether prediction is a useful criterion in evaluating the utility of interim assessments. To do so, we use ordinary least squares and logistic regression to predict scale scores and proficiency classifications on a sixth grade, large-scale summative mathematics assessment using scores from (a) three sixth grade mathematics interim assessments, which were administered quarterly, and (b) a large-scale summative assessment administered in the prior school year (i.e., fifth grade). We show that the increase in variability accounted for, or accuracy in prediction, of the interim assessments over and above the fifth grade summative is not particularly large. Thus we conclude that prediction is a relatively low bar to use in evaluating the utility of interim assessments and that other criteria should be considered.

Methods

Data

Sample. The data used in this study comes from a large district in a south-western state. During the 2014-2015 school year, this district administered quarterly interim assessments in multiple grades and subjects, including sixth grade mathematics, which is the focus of this investigation. Specifically, we use data produced by students enrolled in a general sixth grade mathematics course. These students took three interim assessments over the course of the

academic year, administered at end of quarters one, two and three. On average, each interim was administered after approximately 43 days of instruction (the latest the quarter 1 interim could be administered was after 42 days of instruction, for the quarter 2 interim 51 days and for the quarter 3 interim 38 days). These students also took a large scale, summative assessment aligned to the Common Core State Standards at the end of the year and a similar summative assessment in the preceding grade, i.e., a fifth grade large scale, summative assessment.

While the interim assessments were administered at the district level, not every student took every test, nor did every student complete every item. As we note in the next section, we use total scores in our model (and do not impute scores for students with missing items). Using students who had completed all items on the three interims reduced our sample size from 11,539 to 5,201. Relative to the entire population of 11,539 students, the 5,201 included in our sample were more likely to have lower quarterly grades (e.g., mean grade point average of 2.8 for the full set of students in quarter one vs. 2.29 for the sample), identify as female (49% vs. 51%), identify as an ethic minority (52% vs. 61%), have been identified as student with disabilities (13% vs. 22%) and speak a language other than English in the home (28% vs. 34%). The reason(s) for this disparity are unclear, but one likely explanation is that polices within the district prioritized participation on the interims for students in federal accountability subgroups, perhaps in a similar fashion to potential policies around the end of year, summative assessment.

Measures. Each interim assessment was made up of thirty items that generally align to the instruction of the quarter in which it is administered. Although each test was made up of different set of items the distribution of items by item type (multiple choice and open response) were the same across all three interim assessments. Each interim had 24 multiple choice items and 6 open response items in which students were able to write in their answer. All items were

scored dichotomously. The assessments were developed to district specified blueprints by a regional vendor. Given their less than high stakes role, the interim's levels of reliability were reasonable – with a Cronbach's alpha of 0.77 for the first interim, 0.81 for the second interim and 0.85 for the third interim.

The fifth and sixth grade summative assessments appeared to be typical of large-scale summative assessments. Each test was developed to measure the state's academic content standards, in this case the Common Core State Standards. The reliabilities of the assessments ranged from 0.91 to 0.93. The assessments included approximately 50 items per form. Scale scores were created using item response theory methods and achievement levels were set by state educators using the bookmark method.

Analytic Models

We use OLS linear regression to predict sixth grade summative assessment scale scores using the interim and fifth grade summative scores. Similarly, we use logistic regression to predict student proficiency classifications (0= below proficient, 1=proficient or above). This examination of proficiency classifications is particularly relevant, as it is these classifications that are used in current systems of school accountability. All of the models were implemented using R version 3.2.5 (R Core Team, 2016). To make the ranges comparable between the interim assessment and summative assessments, we normalize the scores for each assessment by subtracting that assessment's mean and dividing by its standard deviation.

For the logistic regressions, we also compute classification accuracy and a pseudo-R^2 statistic, Nagelkerke's R^2, using the rms R package (Harrell, 2016). To calculate the former, we create predicted proficiency classifications based on the logistic model (0 = probability of

proficiency < 0.5, 1 = probability of proficiency ≥ 0.5) and compare to the actual values. While this approach is crude, it does provide a quick indication of how accurate predictions based on the model are.

Results

To provide some context we first present Pearson correlations between the variables, as shown in Table 1, before turning to the regression results. Overall, the scores form each interim correlate similarly with one another, with a mean correlation of 0.72. The correlations between scores from each interim and the sixth grade summative assessment are also similar to one another, with a mean correlation of 0.75. Interestingly, the correlations between the fifth grade summative scores have similar patterns as those based on the interim scores. The correlation between the fifth and sixth grade summative scores is 0.74 and the mean correlation between the fifth grade summative scores and the interim assessment score is 0.68. This indicates that, to a large degree, students are ordered in the same way, regardless of the assessment examined.

Table 1. *Correlations between Assessment Scores*.

	Interim 1 Score	**Interim 2 Score**	**Interim 3 Score**	**G5 Score**	**G6 Score**	**G6 Proficiency**
Interim 1 Score	1.00					
Interim 2 Score	0.69	1.00				
Interim 3 Score	0.68	0.73	1.00			
G5 Score	0.67	0.67	0.70	1.00		
G6 Score	0.71	0.74	0.79	0.74	1.00	
G6 Proficiency	0.60	0.64	0.67	0.60	0.80	1.00

Note: The grey cells indicate correlations with the sixth grade summative assessment. Note that the correlations with proficiency are, by definition, weaker due to the range restriction inherent in creating a proficiency classification.

Table 2 presents the results of the OLS regressions. There are a few things to note. In the models that include all three interim assessments, models (5) and (6), Interim 3 is the strongest

predictor of the sixth grade summative scores. However, for the models in which a single interim was used as the only predictor, the coefficient estimates for the interim scores are similar. This suggests that Interim 3 has more unique variability related to the sixth grade summative scores than Interims 1 or 2. Overall, however, the interim assessments have similarly predictive relationships to the sixth grade summative assessment (R^2 values of 0.50, 0.55 and 0.62, for models with interims 1, 2 and 3 as predictors, respectively). Also, the increase in R^2 from the model in which just the fifth grade summative scores are a predictor, to one with the fifth grade summative and all interim scores is 0.14. Finally, the R^2 value for the model including all the assessment scores, model (5) is the same, at two decimal places, as the value for the model with just the fifth grade summative score and Interim 3.

Table 2. *Results from OLS Regressions (Scores on the Sixth Grade Summative Assessment is the Dependent Variable)*.

Predictor	**Model**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Interim 1 Score	0.68				0.20	0.15	0.40		
Interim 2 Score		0.69			0.28	0.23		0.47	
Interim 3 Score			0.78		0.39	0.33			0.53
G5 Score				0.74		0.20	0.43	0.36	0.36
Intercept	-0.06	-0.09	-0.02	0.00	-0.05	-0.04	-0.03	-0.06	-0.01
R^2	0.50	0.55	0.62	0.55	0.67	0.69	0.61	0.62	0.69

Note: All coefficients are significant at $p<0.05$ or lower.

Similar trends can be observed when looking at the logistic regressions in Table 3, which predict the proficiency classifications from the sixth grade summative assessment instead of the scale scores. The classification accuracies range from 79.10% to 85.48%. The model with the highest classification accuracy, model (5), includes all three interim assessments, followed by the model with all three interim assessments and the fifth grade summative assessment, model (6).

The predictive accuracy of the models with any one interim assessment as a predictor is slightly higher than that of the model with just the fifth grade summative assessment (79.14%, 81.99%, 82.27% vs. 79.10%).

Table 3. *Results from Logistic Regressions (Sixth Grade Summative Assessment Proficiency Classification is the Dependent Variable)*.

Predictor	Model								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Interim 1 Score	1.87				0.72	0.57	1.29		
Interim 2 Score		2.09			1.04	0.92		1.60	
Interim 3 Score			2.32		1.36	1.23			1.81
G5 Score				1.99		0.66	1.31	1.14	1.12
Intercept	-0.88	-1.09	-1.01	-0.90	-1.15	-1.18	-0.98	-1.13	-1.07
AIC	4,133	3,736	3,648	4,222	3,110	3,037	3,690	3,437	3,359
Nagelkerke R^2	0.46	0.53	0.55	0.44	0.64	0.65	0.54	0.59	0.6
Classification Accuracy	79.14%	81.99%	82.27%	79.10%	85.48%	85.44%	82.08%	83.42%	83.65%

Note: All coefficients are significant at $p<0.01$.

Conclusion and Significance

The increase in variability accounted for, or accuracy in prediction, of the interim assessments over and above the fifth grade summative is not particularly large. In general, the fifth grade summative assessment preformed similarly, or very slightly worse, than any interim assessment. In terms of accounting for scale score variability, the fifth grade summative assessment had higher R^2 values than Interims 1 and 2, but not Interim 3. Interim 3 was a stronger predictor of performance on the sixth grade summative assessment, possibly because it was administered closer in time to the summative assessment and/or, potentially, because content on the third interim assessment may be over represented on the summative assessment, relative to the content of the other interims.

Finding that only the third interim assessment, administered very close to the end of the year, provided better prediction than the fifth grade summative assessment suggests that the predictive utility of interim assessments, or at minimum the interim assessments we examined, is a fairly weak justification for their use. That is, the summative assessment from the prior year is as predictive as the first two interims, and the third interim may come too late in the semester to provide actionable information. In addition, large scale, summative assessment results are generally produced by a state's accountability system, whereas interim assessments are often purchased by districts at additional cost.

However, these findings do not suggest that interim assessments have limited utility *overall*. It is quite possible that the interim assessments we examined have great value in the school district. For example, one can imagine teachers using the interim assessments as the basis for conversation in their communities of practice, combining these results with their classroom observations to target instruction and so on. Our point is that motivating the administration of interim assessments based on their predictive utility alone is very likely to be insufficient. Those using interim assessments should seek additional criteria to justify their use. The work for Goertz, Oláh and Riggan (2010), for example, provides a strong basis for additional criteria on which the utility of interim assessments could be judged (see, in particular, their implications chapter, p. 224-243).

Finally, the work presented here could be expanded upon in a number of ways. Future work could involve to re-running the models with additional predictors (including student quarterly grades), accounting for measurement error with the regression models, examining data from students in the accelerated math course (which we excluded here), and, pending the availability of data, examining additional grades and subjects. In addition, future work could also

examine the additional criteria for the evaluation of the utility of interim assessments (like those that could be derived from Goertz, Oláh and Riggan, 2010).

References

Goertz, M. E., Oláh, L. N., Riggan, M. (2010). *From Testing to teaching: The use of interim assessments in classroom instruction* (RR-65). Philadelphia, PA. Consortium for Policy Research in Education.

Harrell, F. E., Jr. (2016). rms: Regression Modeling Strategies. R package version 4.4-2. https://CRAN.R-project.org/package=rms.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.