Kansas Interim Assessment Validity Evidence

Based on the Relationship between Interim and Summative Assessment Scores

Neal Kingston, Wenhao Wang, Angela Broaddus, and Laura Kramer

Center for Educational Testing and Evaluation, Lawrence, Kansas

Paper presented at the annual meeting of the American Psychological Association,

Washington, DC, August 2011.

Correspondence concerning this article should be addressed to Neal Kingston, Center for Educational Testing and Evaluation, University of Kansas, Lawrence, KS 66047.

E-mail: nkingsto@ku.edu

Kansas Interim Assessment Validity Evidence

Based on the Relationship between Interim and Summative Assessment Scores

The primary purposes of an interim assessment system are (1) to identify broad problems with curriculum and/or instruction within a school year and (2) to identify specific students who need significant extra help in mastering the year's curriculum. Both of these purposes might lead to the modification or addition of instructional programs or supports. To these end the correlation of interim assessment results with year-end assessments that schools use as measures of adequate yearly progress is a critical component of any validation effort.

A limited number of previous studies have shown a strong relationship between interim assessment and summative assessment scores. This brief literature review will focus on relationships for mathematics tests.

Brown and Coughlin (2007) reported the correlations between the TerraNova (which was not designed to be an interim assessment but is used as such by some schools) and the Pennsylvania System of School Assessments as ranging from 0.67 to 0.82.

Williams (2008) studied the relationship between the Texas benchmark test system and the Texas Assessment of Knowledge and Skills for reading and math of fourth grade students. Benchmark assessments were administered in October, December, and March and the summative assessment was administered in April. Correlations with the summative assessment, based on samples of over 4,000 students, were .58, .58, and .55 for the scores from the three benchmark administrations.

Underwood (2010) investigated whether the District Benchmark Reading Assessment (DBRA) is a predictor of the reading performance and math performance on Florida Comprehensive Assessment Test (FCAT) for tenth grade students. The correlational analysis was conducted on a sample of 2,263 students. The correlation of FCAT math score on DBRA score is 0.64. Benchmark data were collected in November and summative data were collected in March.

Correlations between Northwest Evaluation Association (NWEA) mathematics Measures of Academic Progress (MAP) scores and Kansas State Assessment scores in grades 3-8 ranged from .57 (grade 6) to .84 (grade 7), with a median correlation of .74 (D. Draper, personal communication, July 15, 2011). The analyses data for both the MAP and Kansas State Assessment were collected in the spring.

Data Source

Data for this study came from 91,716 grade 3-8 mathematics test records that were produced by the Kansas Interim Assessment System during the 2010-11 school year. For the analyses described in this paper, records were removed for any of the following reasons:

- student did not also participate in year-end summative assessment,
- student did not finish all items on each section,
- student tested for so little time (fewer than 10 minutes) that results were considered untrustworthy as a measure of true student performance, or
- student tested for so long a time (more than 60 minutes) that results were considered untrustworthy as a measure of true student performance.

The resulting data set contained 42,866 interim assessment test records from 27,640 unique students. This represents approximately 13% of all Kansas students in grades 3-8. For

analyses of student growth, only 3,589 students who participated in all three interim assessment windows were included to ensure that gain score estimates were not affected by teacher choice of which students would participate in each testing window.

Descriptive Statistics

Table 1 presents the sample size (*n*), mean, and standard deviation of (1) all students who participated in the summative state assessment (Total State), (2) all students who took the interim assessment (Total Interim), and (3) students having all three interim assessment scores (Sample) for grades 3-8. In looking at Table 1, it is important to remember that although each grades' score scale ranges from 0-100, no vertical equating has been performed so there is no appropriate way to compare results within any single column.

Table 1

Summative Assessment Descriptive Statistics

Grade	*n*			Mean			Standard Deviation		
	Total State	Total Interim	Sample	Total State	Total Interim	Sample	Total State	Total Interim	Sample
3	31,275	4,178	442	87.3	87.7	88.4	11.0	10.1	9.5
4	31,767	4,026	575	81.8	82.1	82.9	12.5	11.9	11.4
5	31,760	4,007	523	81.2	80.9	82.2	12.8	12.9	11.3
6	31,894	5,408	644	80.7	81.7	83.5	14.5	13.0	11.1
7	32,198	5,161	761	73.1	73.9	76.5	16.1	14.9	13.9
8	31,631	4,860	644	75.5	75.9	76.7	16.4	15.1	14.5

In Table 1, we can see that the students with complete interim data (column 7) had slightly higher mean scores than typical students at their grade who participated in the interim assessment (column 6), who in turn had higher mean scores than the group of all students who participated in the summative assessment (column 5). The slightly higher mean scores of the Total Interim group compared to the Total State group is especially interesting in that Broadus, Kingston, and Kramer (2011) noted that the interim assessment sample has a higher proportion

of demographic groups that typically score below average. For example, 44-47% of the participants in the interim assessment population (depending on the interim window) received free or reduced lunch, versus 40% of the entire state. While intriguing, this finding is insufficient to conclude that use of the interim assessment leads to improved state summative assessment scores. It might be that teachers who chose to participate with the interim assessment tended to be more willing to try innovative instructional approaches and this served as a causal factor.

Also, at each grade level summative assessment scores of the students who participated in all three interim assessments (last column) had a lower standard deviation than the entire group of students who participated in the interim assessment system (penultimate column), which in turn, in all but 5th grade, was lower than the standard deviation for the entire state (antepenultimate column). There is insufficient evidence to support any particular hypothesis for why this has occurred. Self-selection is one possibility.

Mean Growth

Figure 1 presents the scaled scores by assessment window for those students who participated in all three interim assessment windows and took the summative assessment.

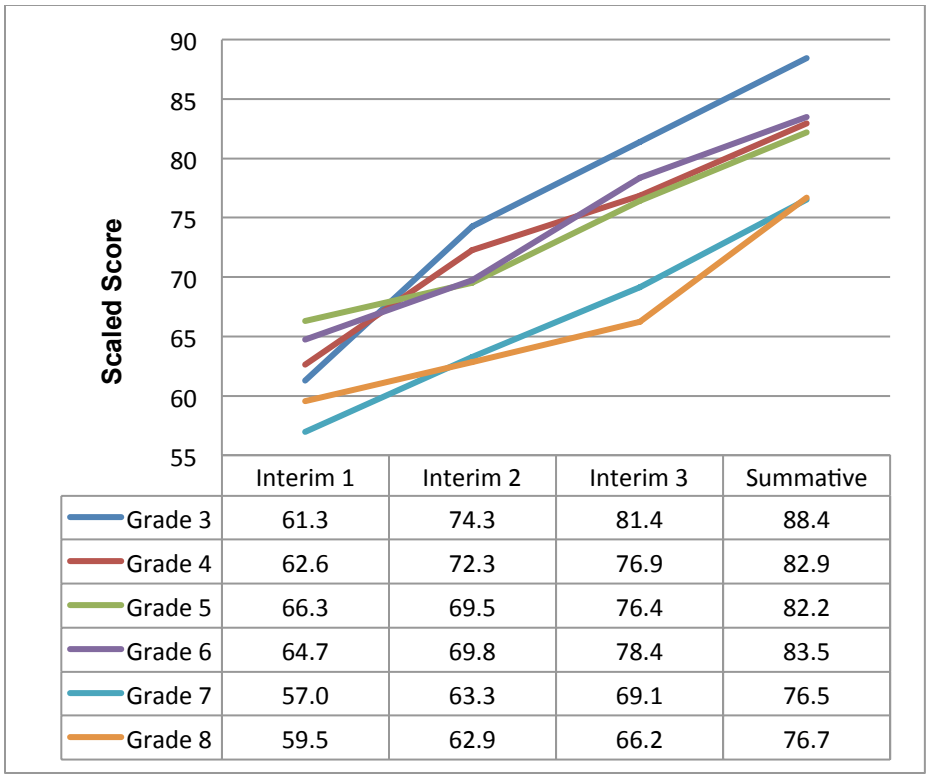

Figure 1: Scaled Scores by Assessment Window.

Table 2 presents the mean and standard deviation of growth for the sample from each interim administration test score to the summative administration score. In addition, it presents the percent of score changes that are greater than or equal to one scaled score point.

Table 2

Growth from Interim Assessment to Summative Assessment

Grade	Mean			Standard Deviation			% Growth >= 1		
	Interim 1	Interim 2	Interim 3	Interim 1	Interim 2	Interim 3	Interim 1	Interim 2	Interim 3
3	27.9	15.0	8.0	24.3	21.8	18.9	98	91	79
4	20.3	10.7	6.1	11.4	9.9	9.6	98	86	71
5	15.9	12.7	5.7	10.0	9.8	8.8	96	90	72
6	18.7	13.7	5.1	10.4	10.4	8.1	97	92	72
7	19.6	13.3	7.4	12.5	11.1	9.8	95	89	77
8	17.1	13.8	10.5	11.1	10.4	10.1	94	92	85

As expected, within each grade level the more instructional time remaining after the interim assessment the more growth occurs. Figure 1shows us that growth is steepest between

the first and second interim assessment windows for grades 3 and grades 4, but that is not the case in the other grades.

Regression Assumptions

While most prediction studies in the social sciences assume a linear relationship between measures of aptitude or achievement it is best to check one's assumptions. This is especially true for this study since issues of motivation could interfere with student performance. While this was not assessed explicitly, we did look at the linearity of regression.

Figure 2 presents a typical scatter diagram using data from the third (winter) grade 8 interim window. A linear and third degree polynomial regression are plotted on the diagram. The lines are very similar for most of the range of data and the r-squared only increased from .6139 to .6164, a difference that is not statistically significant.

While the assumption of heteroscedasticity was not explicitly tested, all scatter diagrams were visually inspected and none appeared problematic. Residual plots were also inspected and deemed to not be problematic.

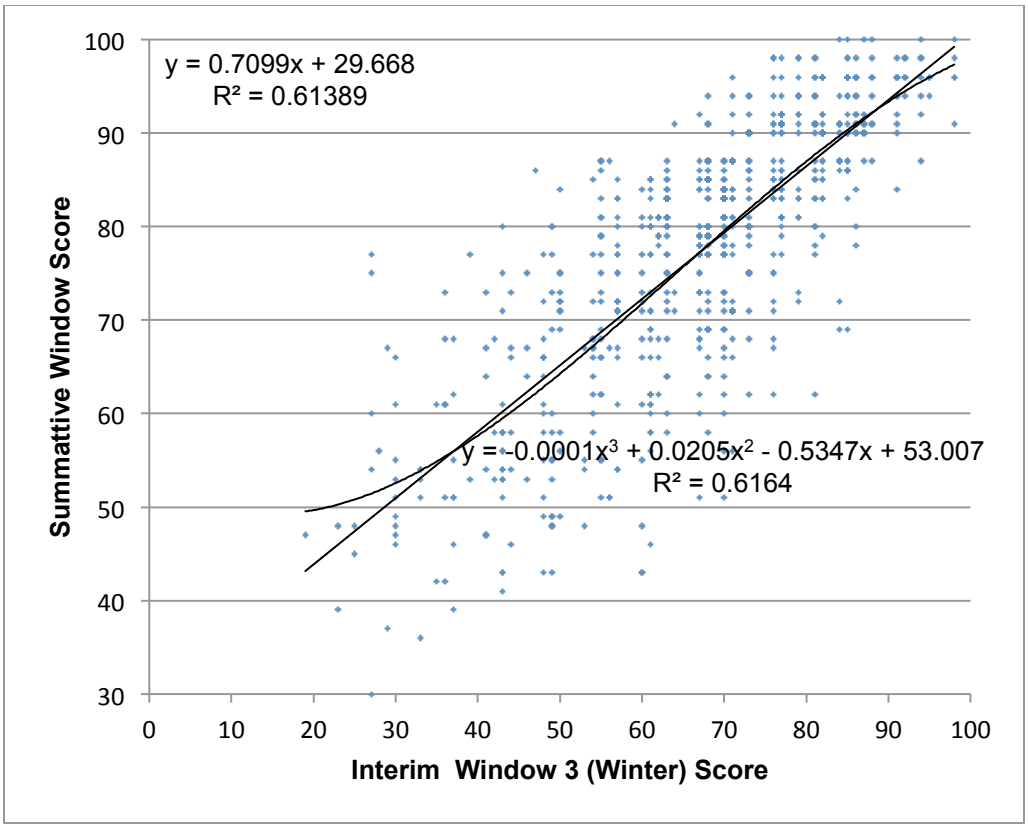

Figure 2: Comparison of Linear and Third Degree Polynomial Regressions. This scatter diagram uses data from the third (winter) grade 8 interim window.

Correlations

Table 3 shows the correlations between the scores from each interim administration and the summative administration.

Table 3

Correlations Between Interim and Summative Scores

Grade	Interim 1	Interim 2	Interim 3
3	.61	.71	.70
4	.65	.72	.73
5	.72	.73	.75
6	.65	.70	.75
7	.70	.74	.80
8	.74	.77	.78

With one exception, the correlation between interim and summative scores is higher for administrations temporally closer to the summative administration. That is, the correlation for Interim 3 is higher than that for Interim 2, which in turn is higher than that for Interim 1. The exception is for grade 3 where Interim 3 correlates with the summative scores .01 less than does Interim 2.

Table 4 is an expansion of Table 3 that also includes multiple correlations for predicting summative test scores from the combination of Interim 1 and Interim 2 scores and the combination of Interim 1, 2, and 3 scores (in the fourth and sixth columns, respectively).

Table 4

Correlations and Multiple Correlations Between Interim and Summative Scores

Grade	Interim 1	Interim 2	Interim 1,2	Interim 3	Interim 1,2,3
3	.61	.71	.72	.70	.76
4	.65	.72	.74	.73	.78
5	.72	.73	.78	.75	.81
6	.65	.70	.73	.75	.79
7	.70	.74	.76	.80	.83
8	.74	.77	.79	.78	.83
Median	.68	.74	.75	.75	.80

From Table 4, we can see that the prediction of summative scores improve when scores from both interim windows 1 and 2 compared to only using the scores from window 2. Based on the way multiple regression works, this must always be the case in the validation sample. However, the correlations remain higher when corrected for shrinkage (that is, when statistically corrected to estimate what the multiple-correlation would be in a cross-validation sample). Moreover, the prediction of summative scores from the first two interim scores is better than the third interim score for four of the six grades (comparing columns four and five). This is a useful finding in that by the third interim window there is little time available for a within-year programmatic response to interim scores. Nonetheless, the improved prediction of summative

scores from all three interim scores, as shown in the last column of Table 4, is of practical as well as statistical significance. For example, the increase in correlation from .70 to .76 is reflected in about a 10% reduction in the standard error of the predicted score.

Regression

Table 5 shows the linear regression equations for predicting summative scores from the interim scores of each administration as well as the multiple regression equation for predicting summative scores from interim scores from the first two and all three windows. A subscript of zero indicates the intercept and all other subscripts indicate the slope for scores from the subscripted administration window.

Table 5

Regression Equations for Predicting Summative Scores

Grade	Interim Window 1		Interim Window 2					Interim Window 3					
	Bivariate		Bivariate		Multivariate			Bivariate		Multivariate			
	b_0	b_1	b_0	b_2	b_0	b_1	b_2	b_0	b_3	b_0	b_1	b_2	b_3
3	69.1	.315	57.7	.414	57.7	.101	.331	45.7	.526	47.2	.042	.227	.268
4	51.7	.499	41.1	.579	39.1	.212	.422	37.0	.597	32.4	.128	.248	.321
5	44.5	.568	41.9	.580	37.9	.307	.344	32.8	.646	30.2	.213	.204	.310
6	48.7	.538	46.3	.533	41.7	.246	.370	28.4	.703	27.7	.136	.175	.442
7	44.8	.556	36.8	.628	36.1	.244	.420	29.2	.684	26.8	.115	.119	.442
8	36.9	.667	32.5	.703	30.0	.316	.442	29.7	.710	24.5	.171	.269	.379

Discussion

Correlations between Kansas Interim Assessment scores and summative test scores were about the same as for other professionally developed tests (for example, a median correlation of .75 compared to a median correlation of .74 for MAP). Not unexpectedly, correlations that are closer in time tend to be greater than those further apart in time. Also not unexpectedly, correlations based on all three interim administrations are meaningfully higher than those based on only one or two administrations.

Correlations may have been limited by the easiness of the tests and a ceiling effect due to the score scale which limited the ability of both the interim and summative assessments to differentiate among relatively high proficiency students.

Recommendations and Future Research

The slightly higher than state average performance of students who participated in the interim assessment is interesting and worthy of follow-up analyses. Among other possibilities, we could look at the previous year summative scores of students who did and did not participate in the interim assessment program.

Some of the impact of ceiling effect could be reduced by using theta estimated rather than scaled scores. This might increase the predictive ability of the interim assessments.

These analyses were based primarily on the students who took all three interim assessments and thus the sample is limited in size. We expect greater use of the interim assessments in 2011-12 and expect that these and other analyses should be replicated.

References

Broadus, A., Kingston, N.M., & Kramer, L. (August 2011). *Implementation Issues and Implications for the Kansas Interim Assessment Program.* Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Brown, R. S., & E. Coughlin. (2007). The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region (Issues & Answers Report, REL 2007–No. 017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid- Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs

Underwood, M. K. (2010). *The Relationship of 10th-grade District Progress Monitoring Assessment Scores to Florida Comprehensive Assessment Test Scores in Reading and Mathematics for 2008-2009.* (Doctoral dissertation, University of Central Florida) Retrieved from http://etd.fcla.edu/CF/CFE0003214/Underwood_Marilyn_K_201008_EdD.pdf.

Williams, L. (2009). *Benchmark testing and success on the Texas Assessment of Knowledge and Skills: A correlational analysis* (Doctorate dissertation, Publication No. AAT 3353754). Phoenix, AZ: University of Phoenix. Retrieved from: http://gradworks.umi. com/33/53/3353754.html.