Statistical and Substantive Comparability:

The Need for Between-Mode Equating And An Equal Number of Test Forms

Submitted to the

[redacted]

STATE X Assessment Advisory Council,
STATE X Technical Advisory Committee,
And
STATE X State Department of Education

December 21, 2005

By

Stephen C. Court
Director of Pupil Evaluation and Testing
Quality Improvement Services

[redacted]

Executive Summary

This paper addresses two recent announcements by the STATE X State Department of Education (STATE X) that affect the comparability of the paper-and-pencil and the computer-based versions (KCA) of the STATE X state assessments.

On December 1, STATE X announced that only one P&P form but multiple KCA forms of the assessment will be available in 2006. On December 5, STATE X announced that no comparability data will be collected in 2006.

The paper discusses the implications of the two announcements and presents empirical evidence to support its findings.

It then offers two recommendations regarding (1) substantive comparability and (2) statistical comparability:

(1) Provide the same number of P&P forms as there are KCA forms.

(2) Equate the KCA and the P&P.

The paper concludes with a call for the state's Technical Assistance Committee and the STATE X Assessment Advisory Council, as well as the leaders of STATE X school districts, to lobby the state to enact these recommendations.

Introduction

At the December 1, 2005 meeting of the STATE X Assessment Advisory Council, Deputy Commissioner of Education, [redacted] divulged that the new generation of the STATE X state math and reading assessments will consist in 2006 of at least four computer-based test forms but only one paper-and-pencil test form. A few days later, on December 5, STATE X confirmed that no comparability data will be collected during the 2006 assessment administration.

The news was a shock. Since January of 2003, the STATE X State Department of Education (STATE X) has been promising districts that equity between the computer-based (KCA) versions and the traditional paper-and-pencil (P&P) versions of the STATE X state assessments would be maintained. Equity is essential because not all schools in all districts possess the technology to test all of their students via computer. Some schools will have no choice for the foreseeable future but to administer the paper-and-pencil versions of the state assessments to at least some if not all of their students. Providing an equal number of identical test forms in both modes has therefore been the first fundamental requisite of maintaining an equitable and substantively comparable dual-mode system – one that will not benefit some schools and districts and disadvantage others. An equal number of KCA and P&P test forms is essential to maintaining the overall interchangeability of the state assessment results.

The second fundamental requisite of maintaining an equitable state assessment system has been statistical comparability. The scores, performance level classifications, and proficiency rates yielded by the two testing modes must be interchangeable at all levels – student, school, disaggregated subgroup, and district. Whether a student is tested via P&P or KCA must not affect that student's state assessment results. In turn, a school's AYP status or its chances of making Standard of Excellence must not depend on whether all its students are tested via KCA, all its students are tested via P&P, or some students are tested via KCA while others are tested via P&P.

This paper first addresses the issue of statistical comparability. It presents evidence from a local study that the two modes do not yield interchangeable results. It then calls for a rigorous, statewide comparability study to determine once and for all whether the KCA and the P&P are comparable and, if indicated, to equate the two modes.

The paper then addresses the issue of substantive comparability. Despite discussion a year ago of moving the KCA from fixed-form to computer-adaptive testing, the substantive comparability of the KCA and P&P has never before been truly at risk. The recent announcement that only one P&P test form per assessment will be provided in 2006 has changed that.

PART 1: Statistical Comparability

According to *Guidelines for Computer-Based Tests and Interpretations* (American Psychological Association [APA], 1986), score comparability or equivalence between computer-based tests and paper-based tests is defined as follows:

> "Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode." (p. 18)

The APA definition naturally pertains not only to test scores, *per se.* It extends, *de facto,* to the performance levels and dichotomized proficiency rates derived therefrom, which are respectively used to classify students and to render decisions about school effectiveness. AYP does not directly represent average test scores but, rather, the percent of students above a given cut point. AYP does not gauge the distance between a school mean and the cut score that delineates proficient from non-proficient. Rather, AYP takes into account the dispersion of individual scores around the mean, as well as the shape of the distribution. Two schools with exactly the same average score can have very different proficiency rates. Similarly, a single assessment administered in two different modes may manifest similar means but still yield very different rates of proficiency. For most instructional and accountability-related purposes, not the scores but the classifications and proficiency rates are what matter, for they are what serve as the basis for decisions and consequences.

The State's Comparability Study

For years, the state's assessment contractor has contended that the KCA and the P&P versions of the tests are statistically comparable. The contention is based solely on the results of a KCA field test conducted in 2003, which yielded data that then were used to study the comparability of KCA and P&P. The comparability study involved a convenience sample of 617 grade 7 students in twelve schools who had taken one form of the math assessment via P&P and a parallel form via KCA. The sample of "double-tested" students was not representative of the state student population. The study involved only six of the state's 304 school districts. The participating districts were small and either rural or suburban. None of the larger, more urban and more ethnically diverse districts participated in the study.

The study involved only the regular edition of the grade 7 math assessment, not the modified or any of the other special tests - the pre-reading, the plain English, the listening, or the Spanish translations. Except for 32 learning-disabled students, the study excluded everyone receiving special education services. No students receiving ESOL services were included. Regular education students needing any kind of testing

accommodation were also left out. Nor did the study investigate the KCA-P&P comparability of any other content area assessments – reading, science, or social studies – at any other grade levels.

Further, the state study focused fundamentally on the comparability of the KCA and P&P strictly in terms of average Total scores and differential item functioning (DIF). It did not look at KCA-P&P differences in terms of either the five performance classifications (Unsatisfactory, Basic, Proficient, Advanced, and Exemplary) or the dichotomized proficiency rates (below Proficient versus Proficient or above). As will be demonstrated, much of importance was overlooked.

Clearly, the state study was too limited in scope and sparse in sample to establish that the KCA and the P&P versions of the assessment yield interchangeable scores, classifications, and proficiency rates - even for the grade 7 math assessment, let alone across other grade levels and content areas.

Yet, STATE X wholeheartedly embraced the study's central finding that there were:

> *"...no meaningful or statistical significant differences in the composite test scores attained by the same students on a computerized fixed form assessment and an equated form of that assessment when taken in a traditional paper and pencil format."*
> (Poggio, Glasnapp, Wang, and Poggio, 2005; page 26 at
>
> http://www.bc.edu/research/intasc/jtla/journal/pdf/v3n6_jtla.pdf .)

STATE X authorized that the KCA be put into full operation in 2004. In doing so, the state effectively declared the KCA and the P&P versions of the STATE X state assessments are comparable – without sufficient evidence to warrant the claim. In math, KCA versions of the General and Modified assessments were made available in 2004 for use at grades 4, 7, and 10. In reading, KCA versions of the General and Modified reading assessments were made available for use at grades 5, 8, and 11.

In 2005, KCA testing was further expanded to include not only math and reading but also science and social studies at all tested grade levels and with all types of students. According to the state's assessment contractor, "Last year (2005), KCA testing occurred in approximately 85% of the districts and in about 75% of the state's buildings, and roughly 58 percent of the students were tested." (Poggio and Consolver, November 28, 2005).

Such rapid adoption of the KCA occurred for three reasons. First, the KCA offered the advantage of "instant" results. Second, districts were sanguine and uncritical in their acceptance of the state study's assurance that the KCA and P&P are comparable. Third, as incentive to schools to double-test as many students as possible, the state agreed to count as final data the higher of each student's two scores.[1]

---

[1] Double-testing involved testing a student twice – once in each mode – with parallel forms of the assessment. Ostensibly, the incentive of counting the higher of each student's two scores was intended to encourage school participation in double-testing in

In combination, these three reasons created powerful momentum toward computer-based assessment administration. Only a lack of logistical capacity at the school level prevented full, one hundred percent participation in KCA testing.

The XYZ Study

An independent study of the 2004 results conducted by the state's largest district, raised questions about the validity of what the state's comparability study had concluded (Court, 2006). The XYZ study analyzed the 2004 and 2005 state assessment results of 2,394 XYZ students who double-tested in elementary and middle school math and reading.[2]

XYZ's initial examination of the 2004 data confirmed what the state study had found in the 2003 pilot data – that the differences in average KCA and P&P scores were small (see Exhibit 1).

Exhibit 1

Means, Standard Deviations, and Counts by Subject Area and Grade

Subject	Grade		KCA	P&P	KCA-P&P Difference
Math	4	Mean	**55.51**	**57.12**	**-1.61**
		Std. Deviation	17.485	17.096	9.375
		Number of Students	753	753	753
	7	Mean	**45.36**	**48.08**	**-2.72**
		Std. Deviation	17.432	18.092	9.822
		Number of Students	465	465	465
Reading	5	Mean	**77.65**	**78.39**	**-.73**
		Std. Deviation	10.735	10.753	5.750
		Number of Students	763	763	763
	8	Mean	**76.34**	**77.87**	**-1.53**
		Std. Deviation	10.904	10.462	5.556
		Number of Students	414	414	414

Exhibit 1 shows small mean differences that range from a low of 0.73 points for grade 5 reading to a high of 2.04 points for grade 7 math. In both content areas and at all grade levels, the P&P means were slightly higher than the KCA means, with effect sizes that ranged from .07 to .16.

However, means - and the differences between them - do not reveal the entire picture. The differences between the KCA and P&P scores of the double-tested students were found to run about equally in both

order to increase the amount of double-test data collected. However, the incentive also served to motivate schools to participate in KCA testing, thereby accelerating the momentum toward KCA.

[2] Too few high school students double-tested in Wichita – 13 for grade 10 math and 22 for grade 11 reading – for the Wichita study to examine comparability at the high school level.

directions. Some students earned higher scores in the KCA mode, some students earned higher scores in the P&P mode, and some students even earned KCA and P&P scores that were exactly the same.

Exhibit 2

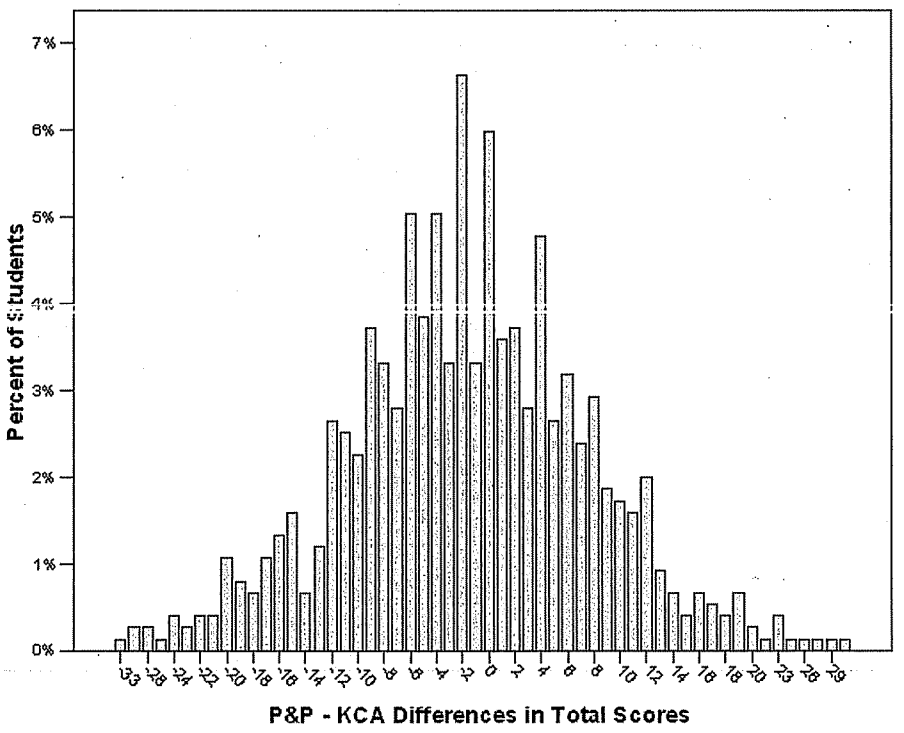
Number and Percent of Students Scoring Higher on the KCA, Higher on the P&P, Or Exactly the Same in Both Modes

Subject	Grade		Number of Students	Percent of Students
Math	4	P&P Higher	415	55%
		No Difference	45	6%
		KCA Higher	293	39%
		Total	753	100%
	7	P&P Higher	258	55%
		No Difference	25	5%
		KCA Higher	182	39%
		Total	465	100%
Reading	5	P&P Higher	388	51%
		No Difference	57	7%
		KCA Higher	318	42%
		Total	763	100%
	8	P&P Higher	221	53%
		No Difference	45	11%
		KCA Higher	148	36%
		Total	414	100%

Using the 2004 grade 4 math results as an example, Exhibit 3 presents a visual display of how the KCA-P&P score differences were distributed.

Exhibit 3

P&P – KCA Differences in Total Scores – Grade 4, 2004



With the distribution being so symmetrical in shape, the negative differences and the positive differences tend to cancel out, yielding a mean difference of merely -1.61 points. That is, the average KCA score, district-wide, was 1.61 points lower than the average P&P score. Nevertheless, at the individual student level, the score differences were as great as 33 points in favor P&P at one extreme and as great as 32 points in favor of KCA at the other extreme.

Focusing exclusively on mean score differences leads KCA-P&P comparability to appear much greater than it actually is. Looking at the absolute values of the score differences begins to clarify the picture (see Exhibit 4).

Exhibit 4

Absolute Differences in P&P – KCA Total Scores

Subject	Grade		Frequency	Percent	Cumulative Percent
Math	4	No Difference	45	6.0%	6.0%
		1 to 5 Points	299	39.7%	45.7%
		6 to 10 Points	220	29.2%	74.9%
		11 to 15 Points	107	14.2%	89.1%
		16 to 20 Points	56	7.4%	96.5%
		More than 20 Points	26	3.5%	100.0%
		Total	753	100.0%	
	7	No Difference	25	5.4%	5.4%
		1 to 5 Points	151	32.5%	37.8%
		6 to 10 Points	137	29.5%	67.3%
		11 to 15 Points	101	21.7%	89.0%
		16 to 20 Points	34	7.3%	96.3%
		More than 20 Points	17	3.7%	100.0%
		Total	465	100.0%	
Reading	5	No Difference	57	7.5%	7.5%
		1 to 5 Points	453	59.4%	66.8%
		6 to 10 Points	204	26.7%	93.6%
		11 to 15 Points	39	5.1%	98.7%
		16 to 20 Points	9	1.2%	99.9%
		More than 20 Points	1	.1%	100.0%
		Total	763	100.0%	
	8	No Difference	45	10.9%	10.9%
		1 to 5 Points	244	58.9%	69.8%
		6 to 10 Points	93	22.5%	92.3%
		11 to 15 Points	25	6.0%	98.3%
		16 to 20 Points	6	1.4%	99.8%
		More than 20 Points	1	.2%	100.0%
		Total	414	100.0%	

At grades 4 and 7, nearly 11 percent of the KCA and P&P math scores differed by more than 15 points. At grades 5 and 8, the extreme differences between KCA and P&P were less pronounced, with less than two percent of the scores differing by more than 15 points. The reading assessments are not necessarily more comparable than the math assessments, however. Rather, the range of possible reading differences is merely more restricted than the range of possible math differences.[3]

[3] This is so because the average KCA reading score and the average P&P reading score were considerably higher than the average math scores – about 80 for reading as compared with about 56 for grade 4 math and 47 for grade 7 math. So, with both the KCA and P&P average reading score being closer to the maximum possible score of 100% correct, the range of score differences was considerably more restricted.

To clarify the picture even more, Exhibit 5 collapses the categories. Panel A shows the number and percent of students whose KCA and P&P scores differed by more than 5 points. Panel B shows the number and percent of students whose KCA and P&P scores differed by more than 10 points.

Exhibit 5

Panel A

Number and Percent of Students with Score Differences Greater than 5 Points

Subject	Grade		Number of Students	Percent of Students
Math	4	0 to 5 Points	344	46%
		More than 5 Points	409	54%
		Total	753	100%
	7	0 to 5 Points	176	38%
		More than 5 Points	289	62%
		Total	465	100%
Reading	5	0 to 5 Points	510	67%
		More than 5 Points	253	33%
		Total	763	100%
	8	0 to 5 Points	289	70%
		More than 5 Points	125	30%
		Total	414	100%

Panel B

Number and Percent of Students with Score Differences Greater than 10 Points

Subject	Grade		Number of Students	Percent of Students
Math	4	0 to 10 Points	564	75%
		More than 10 Points	189	25%
		Total	753	100%
	7	0 to 10 Points	313	67%
		More than 10 Points	152	33%
		Total	465	100%
Reading	5	0 to 10 Points	714	94%
		More than 10 Points	49	6%
		Total	763	100%
	8	0 to 10 Points	382	92%
		More than 10 Points	32	8%
		Total	414	100%

With such large percentages of students exhibiting such large differences between their two scores, the KCA and the P&P do not yield scores that are interchangeable. The two modes are not statistically comparable.

Performance Level Classifications and Proficiency Rates

The state study based its conclusion that the KCA and P&P modes are comparable on the small differences in ***average*** Total scores. The XYZ study found that these small mean differences actually mask much larger differences between performance classifications (see Exhibit 6).

Exhibit 6

Differences in KCA and P&P Performance Classifications

Subj	Grade		Percent of Students	Number of Students
Math	4	No Change	52.2%	393
		1 Level	43.4%	327
		2 Levels	4.1%	31
		3 Levels	.3%	2
		Total	100.0%	753
	7	No Change	49.9%	232
		1 Level	43.9%	204
		2 Levels	6.2%	29
		Total	100.0%	465
Reading	5	No Change	55.0%	419
		1 Level	41.5%	316
		2 Levels	3.4%	26
		3 Levels	.1%	1
		Total	100.0%	762
	8	No Change	58.5%	242
		1 Level	38.6%	160
		2 Levels	2.9%	12
		Total	100.0%	414

For at least 45% of the double-tested students, the KCA and the P&P performance level classifications differed by at least one performance level. The bad news is that such high rates of student differential classification yield ambiguous information that leads to poor instructional and administrative decisions. The good news is that only a subset of these differences affects proficiency rates and AYP status – specifically, the difference in classifications between Basic and Proficient.

The differences in proficiency rates – the percent of students below proficient versus the percent of students proficient or above - between the KCA and P&P assessments are displayed in Exhibit 7.

Exhibit 7

Differences in KCA and P&P Proficiency Rates

Subject	Grade		Percent of Students	Number of Students
Math	4	P&P Proficient Only	10.4%	78
		No Difference	84.2%	634
		KCA Proficient Only	5.4%	41
		Total	100.0%	753
	7	P&P Proficient Only	11.4%	53
		No Difference	83.4%	388
		KCA Proficient Only	5.2%	24
		Total	100.0%	465
Reading	5	P&P Proficient Only	9.4%	72
		No Difference	84.6%	645
		KCA Proficient Only	5.9%	45
		Total	100.0%	762
	8	P&P Proficient Only	10.4%	43
		No Difference	84.8%	351
		KCA Proficient Only	4.8%	20
		Total	100.0%	414

Regardless of grade level and content area, roughly 15% of the double-tested students were classified as proficient in one mode but not the other. About 10% of the students exhibited proficiency on the P&P but not the KCA, compared with 5% of the students exhibiting proficiency on the KCA but not the P&P. Note that preliminary analyses of the 2005 double-test data reveal that the overall difference in KCA and P&P proficiency rates remained at roughly 15% in both math and reading at the district level.

School-level Differences

At the school level, the differences between the KCA and the P&P results tended to be larger than the district-wide differences. Frequently, differences in KCA and P&P scores exceeded 20 percentage points. Presumably, these larger differences are due to smaller N's, as well as to substantive differences between schools in how much technology they possess and how that technology is employed. Consequently, school-level differences tended to be not only larger but also more variable. In fact, KCA proficiency rates exceeded P&P proficiency rates in one-quarter of the XYZ schools that double-tested – specifically, in eleven of forty-four cases across math and reading,.

Administering only the P&P version of the state assessments would not resolve the comparability issue. Not all proficiency rates would be higher if ***all*** schools were to do P&P. Thus, decisions regarding whether to administer the state assessments via KCA or P&P should not be made at the district level or even on a

school-wide basis. Rather, such decisions should occur on a more specific basis: grade level by grade level and content area by content area, or classroom-by-classroom, or – ideally – student-by-student.

Further, while the differences between KCA and P&P results run in both directions for individual students, they do not run in both directions at the school level – at least not when the higher of each student's two scores counts as final. The point is important: except when KCA results are used in an unethical manner to influence subsequent P&P scores, a lack of comparability between KCA and P&P results never leads to an over-estimate of a student's performance level classification or a school's proficiency rate. Student performance and school proficiency rates can only be under-estimated.

At least some of the schools that did not make adequate yearly progress in 2004 or 2005 may very well have made AYP if only they had administered the state assessments to each student in the mode that would have yielded the higher score had the student been double-tested.

Disaggregations

The XYZ study found that the differences in proficiency rates did not vary significantly across demographic disaggregations, except within a small number of certain schools. District-wide, the KCA-P&P differences were similar across racial, gender, and socioeconomic groups, as well as across educational program (special ed, ESOL, etc). The KCA and the P&P assigned about 45% of any disaggregated group to different performance levels. For any subgroup, the KCA and P&P proficiency rates differed by roughly 15%, regardless of grade level and content area. Exhibit 8 displays the differences in performance level classifications across educational program for the reading assessment.

Exhibit 8
Differences in Performance Level Classifications across Educational Program

Subj	Grade			No Change	1 Level	2 Levels	3 Levels	Total
Reading	5	Reg	Count	314	259	19	1	593
			Row %	53.0%	43.7%	3.2%	.2%	100.0%
		ELL	Count	59	35	4	0	98
			Row %	60.2%	35.7%	4.1%	.0%	100.0%
		SPED	Count	46	22	3	0	71
			Row %	64.8%	31.0%	4.2%	.0%	100.0%
		Total	Count	419	316	26	1	762
			Row %	55.0%	41.5%	3.4%	.1%	100.0%
	8	Reg	Count	197	136	9		342
			Row %	57.6%	39.8%	2.6%		100.0%
		ELL	Count	17	9	2		28
			Row %	60.7%	32.1%	7.1%		100.0%
		SPED	Count	28	15	1		44
			Row %	63.6%	34.1%	2.3%		100.0%
		Total	Count	242	160	12		414
			Row %	58.5%	38.6%	2.9%		100.0%

Exhibit 9 displays the degree to which KCA and P&P proficiency rates differ by gender for the math assessment.

Exhibit 9
Differences in Proficiency Rates by Gender

Subject	Grade			KCA and P&P agree	KCA and P&P disagree	Total
Math	4	Female	Count	333	61	394
			Row %	84.5%	15.5%	100.0%
		Male	Count	300	58	358
			Row %	83.8%	16.2%	100.0%
		Total	Count	633	119	752
			Row %	84.2%	15.8%	100.0%
	7	Female	Count	204	37	241
			Row %	84.6%	15.4%	100.0%
		Male	Count	184	43	227
			Row %	81.1%	18.9%	100.0%
		Total	Count	388	80	468
			Row %	82.9%	17.1%	100.0%

Mode and the DT Effects

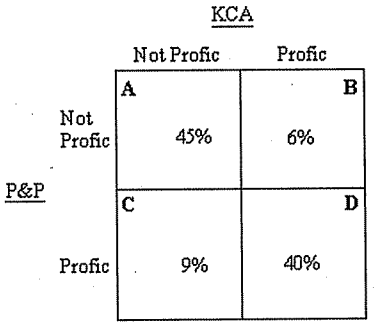
The XYZ study found that differences between KCA and P&P results can be partitioned into two components: definite and provisional. A definite classification occurs when both the KCA and the P&P

results agree that a student is proficient or not proficient. A provisional classification occurs when the KCA and the P&P results disagree – that is, when a student is found to be proficient by one mode but not the other. Provisional proficiency represents the amount by which the proficiency rate would increase if every student were either (a) double-tested and awarded his or her higher score or (b) tested just once in the mode that would elicit his or her optimal performance. When the higher of each student's two scores counts, final proficiency is the sum of definite proficiency and provisional proficiency.

In turn, the provisional proficiency rate can be partitioned into a "mode effect" and a "double-test" (DT) effect. Exhibit 10 provides a hypothetical but realistic example.

Exhibit 10

Example of Mode Effect and DT Effect



Definite Proficiency = Cell D = 40%

Provisional Proficiency = Cell B + Cell C = 6 + 9 = 15%

Final Proficiency (if each student's higher score counts) = Definite Proficiency + Provisional Proficiency = 40 + 15 = 55%.

Mode effect = absolute value of Cell B – Cell C = absolute value of 6 - 9 = 3%.

DT effect = Provisional Proficiency – Mode effect = 15 - 3 = 12%.

Partitioning provisional proficiency into distinct mode and DT effects is important to understanding how different levels of equating might affect final proficiency. In the example provided in Exhibit 10, fifteen percent of the students were provisionally proficient. If every student were either double-tested and awarded his or her higher score or tested just once in the mode that would elicit his or her optimal performance, the 15% provisional proficiency rate would count with the 40% definite proficiency rate and the final proficiency rate would be 55%.

In turn, the 15% provisional rate partitions into a 3% mode effect and a 12% DT effect. If equating adjusted only for the mode effect, the final proficiency rate would increase merely to 43%, leaving school effectiveness under-estimated. Only if equating adjusted for both the mode effect and the DT effect would the final proficiency rate increase to 55%, thus representing amore accurate estimate of school effectiveness..

The Mode Effect and the DT Effect - Conceptually

What the research literature refers to technically as a "mode effect" represents the average differences in KCA and P&P results at the school, district, or state level. Whether in reference to scores or proficiency rates, the mode effect is simply the difference between the overall KCA average and the overall P&P average – without regard for the average difference between the higher and lower scores of individual students.

The mode effect represents a "between-subjects" or an "independent groups" effect, as if one group of students had taken the KCA and an entirely different but similar group of students had taken the P&P. As far as the mode effect is concerned, no student was tested twice, and so no student was awarded the higher of his or her two scores. The mode effect does not consider within-student variation.

But for two years, thousands of STATE X students actually ***were*** double-tested, the higher of their two scores ***did*** count, and the proficiency rates and AYP status of the schools that double-tested have been based on the higher of each student's two scores. This is what the XYZ study refers to as the "DT effect." The DT effect represents the average "within-student" differences in scores, performance level classifications, or proficiency rates - over and above the mode effect. Alternatively, one can conceptualize the DT effect as student-by-mode interaction. In any case, the DT effect is more than mere residual or random error.

Throughout STATE X for the last two years, the DT effect has impacted state assessment results - at least those results associated with double-tested students. Double-testing and counting each student's higher score has served as an *ad hoc* equating adjustment that has compensated those schools that double-tested for their lack of omniscience in knowing *a priori* whether KCA or P&P would elicit a particular student's optimal performance.

Thus, equating to the DT effect, as well as the mode effect, would serve as a school-level accommodation for all schools, now that the practice of double-testing has been discontinued. Unless the KCA and the P&P results are equated, schools that have double-tested in previous years will almost certainly experience

lower proficiency rates.[4] More importantly, unless the KCA and the P&P results are equated, the proficiency rates of schools that did not double-tested will continue to be underestimated – to a degree that would prevent AYP determinations from being valid indicators of school effectiveness.[5]

Unlike the mode effect, the double-test effect would not register in a random equivalent groups design. The DT effect is detectable only with a repeated-measures research design. That is why, if ever the KCA and the P&P are to be equated, a repeated measures research design must be employed in future comparability studies: so that there will be double-test data to adjust not only for the average KCA-P&P differences that constitute the Mode effect but also for the within-student differences that constitute the DT effect.

Double-test Legitimacy

The legitimacy of the double-test effect has been the most disputed aspect of XYZ's comparability study. The controversy involves whether the higher of a student's two scores necessarily represents the more valid score. Some experts contend that it does. If a student had trouble scrolling through the passage when taking the KCA version of the assessment, for instance, then it is likely that the P&P version elicited a score not only higher but also closer to the student's true score. Because scrolling skill is irrelevant to reading comprehension, the P&P score would necessarily be the more valid reflection of the student's reading comprehension. Conversely, if a student is more engaged and thus more motivated when taking a computer-based test, then the KCA would be the more valid score.

However, other experts argue that if the higher of the two scores is associated with the second testing occasion, then it likely is due either to a practice effect from having taken the assessment twice or, more nefariously, to over-zealous teachers using the "instant" KCA results as a basis for instructional intervention before their students take the P&P version – in short, to cheating.

The key phrase in the previous paragraph is "if the higher of the two scores is associated with the second testing occasion." To resolve this dispute, XYZ did keep track of the sequence in which each of its double-tested students took the two tests in 2005. Analysis of the sequence data shows clearly that practice effects

---

4 The decline in proficiency rates of schools that double-tested in previous years will be tempered by two factors: (1) the degree to which they have improved their curricular and instructional effectiveness in order to yield overall increases in student performance and (2) the degree to which they accurately predict whether particular students will perform better on the KCA or the P&P version of an assessment. The first factor, of course, is perfectly relevant to AYP, while the second factor is perfectly irrelevant to AYP.

5 Just as providing an assessment in Braille to a visually-impaired student offsets the student's inability to see, equating to both the mode effect and the DT effect offsets a school's lack of omniscience. Just as testing a visually-impaired student without a Braille or listening accommodation would not elicit scores that serve as a valid indicator of what the student actually knows and can do, not equating the KCA and the P&P results would not elicit proficiency rates that serve as accurate indicators of school effectiveness.

were negligible. Using the 2005 grade 8 reading results as an example, Exhibit 11 displays the proficiency rates broken out by test administration sequence.

Exhibit 11

Proficiency Rates by Test Administration Sequence (Grade 8 Reading)

Sequence		Mean	N	Std. Deviation	Std. Error of Mean
P&P First	P&P Score	80.1	376	10.03	.52
	KCA Score	79.0	376	9.73	.50
	P&P Proficiency	60.4%	376	48.98	2.53
	KCA Proficiency	55.6%	376	49.75	2.57
KCA First	P&P Score	81.9	519	9.84	.43
	KCA Score	80.5	519	10.57	.46
	P&P Proficiency	64.2%	519	48.00	2.11
	KCA Proficiency	60.9%	519	48.85	2.14

Regardless of whether students were first administered the KCA or the P&P version of the assessment, the P&P averages were higher than the KCA averages, but the differences were small and consistent across sequence. The average P&P score of students who took P&P first was 1.1 points higher than the average KCA score. In turn, among the P&P-first students, the proficiency rate was 4.8% higher than the KCA proficiency rate. In comparison, among the students who took KCA first, the average P&P score was 1.4 points higher than their KCA score, and the P&P proficiency rate was 3.3% higher than the KCA proficiency rate.

What matters with respect to identifying the presence of a practice effect, though, are the differences of the differences. A significant practice effect would have been indicated by the difference of the KCA-first proficiency rates being much larger than the difference of the P&P-first proficiency rates.

But, analysis revealed that the gaps in scores and proficiency rates were statistically similar. As shown in Exhibit 12, the differences between the KCA-first differences and the P&P-first differences were not large enough to be statistically significant.

Exhibit 12

Significance Tests of the Score and Proficiency Differences between Taking P&P or KCA First

	Sequence	N	Mean Difference (P&P - KCA)	Std. Deviation	Std. Error Mean
Score	P&P First	376	1.125	5.438	.280
	KCA First	519	1.364	5.476	.240
Proficiency	P&P First	376	4.787	41.678	2.149
	KCA First	519	3.276	36.350	1.596

Independent Samples Test

			Levene's Test for Equality of Variances		t-test for Equality of Means		
		Mean Difference	F	Sig.	t	df	Sig. (2-tailed)
Score	Equal variances assumed	-.239	.811	.368	-.647	893	.52
	Equal variances not assumed	-.239			-.647	811.3	.52
Proficiency	Equal variances assumed	1.512	5.185	.023	.577	893	.56
	Equal variances not assumed	1.512			.565	739.6	.57

The two critical pieces of information in Exhibit 12 are circled. For the tests to be "positive" – that is, to indicate that student performance was influenced by a practice effect – the circled values would need to be ".05" or smaller. They are not. Taking the test twice did not significantly affect the differences in the scores or the proficiency rates between the KCA and the P&P.

Why is this important? It is important because it supports the contention that the DT effect is real and legitimate – the higher of each double-tested student's two score is the more valid score.

So, why is that important? That is important because it supports the argument that the KCA and the P&P should be equated to adjust for not only the mode effect but also the DT effect – for not only double-tested students but also single-tested students. In other words, the equating should adjust for the average difference between KCA and P&P scores or between the KCA and P&P proficiency rates. Moreover, it should also adjust for how well each student would have scored if he or she had taken the assessment in the mode that would have yielded his or her better performance.

Of course, we cannot know with absolute certainty how a single-tested student would score in another mode. But tried-and-true statistical techniques do exist for estimating with reasonable precision what a student's KCA score would be from his or her P&P score, or vice-versa. Such estimation would be based

on the ranges of KCA scores observed in the double-test data that correspond to each particular P&P score or, conversely, the ranges of P&P scores observed in the double-test data that correspond to each particular KCA score. After all, that was the motivation for collecting double-test data in the first place – to provide a basis for equating the two modes if they proved not to be comparable – not just for double-tested students but for all test-takers.

But, what if such equating studies had been conducted – how much difference might equating have made?

The XYZ comparability study used the 2004 double-test data to perform impromptu equipercentile and linear equating adjustments to simulate the AYP status of all XYZ schools, regardless of whether or not they participated in the double-testing. Exhibit 13 displays these results, which do not include the application of confidence intervals or Safe Harbor.

Exhibit 13

If KCA and P&P had been equated in 2004 for both the Mode and the DT effects, the following schools which did not actually make adequate yearly progress would have made AYP.

	School	Number of subgroups whose AYP status would have improved
	East H.S.	1
	North H.S.	3
	South H.S.	4
	Southeast H.S.	4
	West H.S.	3
	Heights H.S.	3
	Brooks Middle	1
	Coleman Middle	1
	Curtis Middle	1
	Hadley Middle	1
	Jardine Middle	1
	Mead Middle	1
	Stucky Middle	1
	P.V. Middle	1
	Cloud Elem.	1
	Colvin Elem.	1
Total	16	28

No fewer than 28 disaggregated subgroups within 16 schools that did not make AYP in 2004 would have made AYP if the KCA and P&P had been equated.

Although a formal simulation with the 2005 data has not yet been completed, preliminary analyses of the data suggest that the results for 2005 would be similar.

The Technical Advisory Committee – composed of Professors Popham, Linn, Kolen, Pelligrino, and Thurlow - has twice recommended that STATE X conduct such studies. In its peer-review guidance (OESE, 2004), the U.S. Department of Education has made such studies a criterion in evaluating the technical quality of state assessments and state accountability systems (see section 4.4). Several renowned assessment theorists, researchers, and practitioners, as well as all the major sets of professional assessment standards, strongly recommend that such studies be conducted (see Paek, P., 2005; Wang and Kolen, 2001, as well as AERA, 1999; APA, 1986).

Yet, the double-test data collected for the last two years have not been used by the state to conduct studies to investigate the comparability of the two modes and, if indicated, to equate the KCA and the P&P even for the Mode effect, let alone for the DT effect.

With respect to the future - for several months, on several occasions and in public venues, the state has promised that 2006 would be the year when a rigorous and unassailable study would finally be conducted to confirm the comparability of the KCA and the P&P or, if indicated, to equate the two modes equated. (See http://www.STATE X.org/outcomes/readassesssum.doc, page 2, under "Computerized Assessments" – last updated 8-2-05; or see Randall, September 26, 2005 at http://www.STATE X.org/assessment/assessguidelines.ppt#276,27,DoubleTesting).

Yet, on December 5, 2005, the state divulged that it has decided that no comparability data will be collected during the 2006 assessment administration. Despite the recommendations of the TAC and the U.S. Department of Education, no "scientific" study, employing a rigorous experimental design with counterbalancing and accounting of administration sequence will be conducted. The explanation provided by STATE X is that "We will wait and see how the Feds respond to our peer review material" (C. Randall, personal correspondence.)

Near the beginning of this paper, it was stated: "*...A school's AYP status or its chances of making Standard of Excellence should not depend on whether all students are tested via KCA, all students are tested via P&P, or some students are tested via KCA while others are tested via P&P.*"

The XYZ study clearly shows that both AYP status and the chances that a school will make Standard of Excellence do depend on the mode in which each student is tested. Without proper equating or without a perfect method of predicting which mode will elicit each student's more optimal and valid performance, the proficiency rates of schools have been - and likely will continue to be - underestimated.

This holds not only for schools that double-tested but even moreso for schools that opted to do KCA only or P&P only. Not only in XYZ but across the entire state, thousands of schools have assumed that their state assessment results are pure reflections of what their students know and can do, that the mode of testing makes no difference, and that the new generation of tests will be even better than the last.

That is why the news that only one form of the P&P would be available in 2006 was so shocking. With only one P&P form but four or more KCA forms, the differences in scores and proficiency rates will probably increase, as will be demonstrated below. Further, with only one P&P form but four or more KCA forms, the proper and accurate equating of the KCA and P&P cannot possibly occur – even if the state were inclined to conduct such a study. This is so because equating can be done only when tests measure the same thing and with equal precision (Porter, 1991; Linn and Kiplinger, 1994). As will be shown in a moment, the difference in content coverage and measurement precision due to the disparity in the number of test items employed by the KCA and the P&P in 2006 now casts considerable doubt upon whether the two modes of testing will yield results that could be equated.

PART II: Substantive Comparability

As far back as December 18, 2003, STATE X had formally approved two resolutions by the STATE X Assessment Advisory Council's that:

- "schools must receive individual student reports of assessed indicators," and
- "there will be the same number of versions of the paper/pencil and computer-based tests available through 2006-2011 testing cycle"

The second resolution was issued to ensure that the indicator reports would be of equal quality regardless of whether they were based on KCA or P&P data.

STATE X's recent decision to provide an unequal number of P&P and KCA test forms both contravenes the agreement and, more importantly, jeopardizes the substantive comparability of the dual-mode state assessment system.

Using one paper-and-pencil form (P&P) but four or more KCA forms jeopardizes substantive comparability by impacting the state assessment results in three ways: (1) it will impact content coverage and thus construct validity (2) it will impact precision and thus reliability; and (3) it will impact scores and thus performance level classifications and proficiency rates.

Impact on Content Coverage and Construct Validity

When combined to make school-level or district-level decision, the four or more KCA forms will provide greater content coverage than will the one P&P form in terms of depth, breadth, or both. The four KCA forms will provide much "deeper" content coverage to the degree that their items are "clones" of each other (see Exhibit 14).[6]

Conversely, to the degree that their items vary in what or how they measure different indicators, or in their levels difficulty and discrimination, the four KCA forms will provide much "broader" content coverage.

---

6 "Clones" are test items that measure particular indicators in exactly the same way and with the same level of difficulty and discrimination.

Exhibit 14

Simple Examples of "Cloned" and "Not Cloned" Items

Clones		Not Clones	
Item A	Item B	Item C	Item D
5+3=	6+2=	6+2=	6x2=
A. 2	A. 4	A. 4	A. 3
B. 8	B. 8	B. 8	B. 4
C. 35	C. 26	C. 26	C. 8
D. 53	D. 62	D. 62	D. 12
		Item E	
		Mary has 6 marbles. She gets 2 more. How many marbles does Mary have now?	
		A. 4	
		B. 8	
		C. 26	
		D. 62	

Note: Item A and Item B are clones because they measure exactly the same skill, in exactly the same way, with exactly the same kind of item. Although Item C and Item D use the same kind of item, they are not clones because they measure different skills (addition and multiplication) while Item C and E are not clones because they use measure the same skill with different item types.

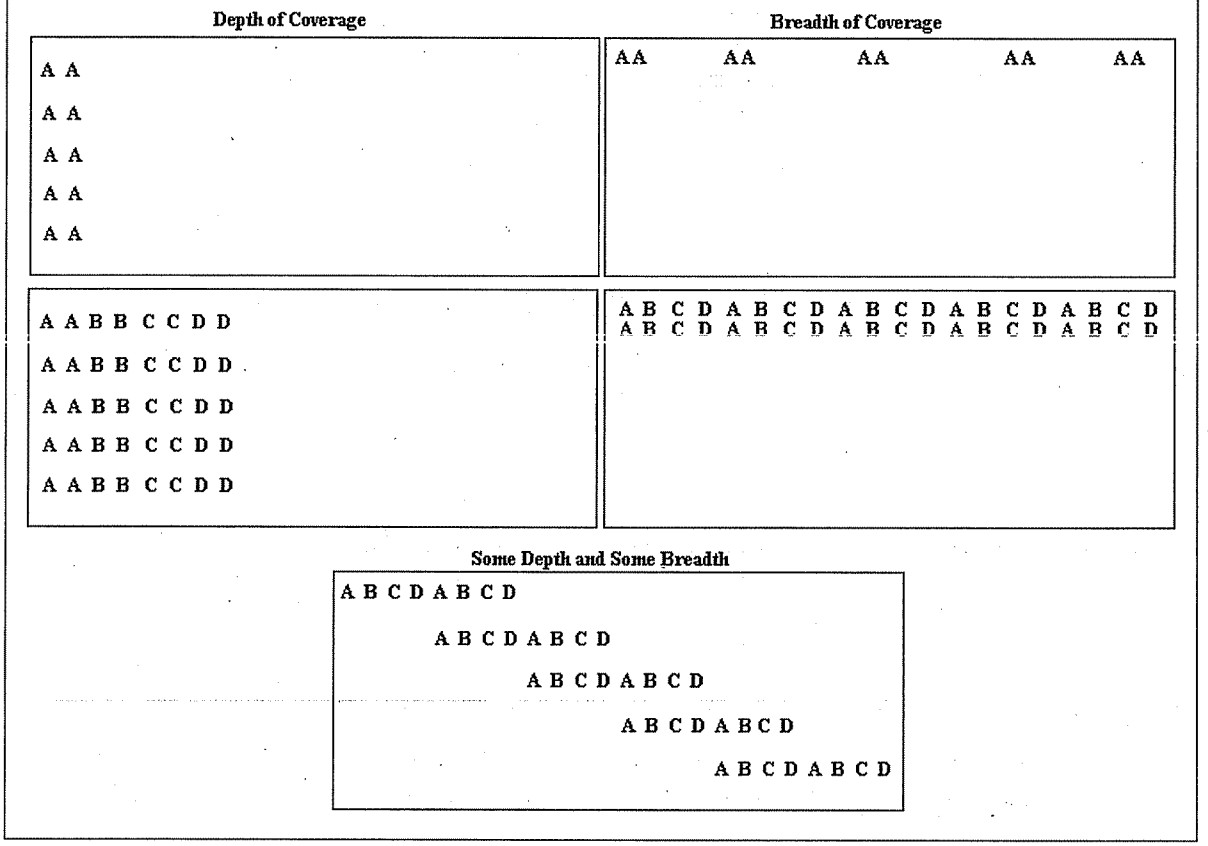
"Deeper" content coverage is associated with greater reliability. "Broader" content coverage is associated with greater construct representation and, therefore, with greater validity. There usually is a trade-off between validity and reliability. The trade-off is remedied through increasing the length of the test by adding more test items.

Four KCA forms will mean that school-level assessment results will be based on four times as many items, which will tend to balance out the trade-off between validity and reliability at an acceptable level. This will not be the case with only one form of the P&P.

Exhibit 15 provides visual displays of deep and broad content coverage. The example assumes a unidimensional construct – for example, "Math Computation" – measured by 10 items per form.

Exhibit 15

Visual Displays of "Deep" and "Broad" Content Coverage



Given that any individual student will take only one form of the test, the number of forms will not affect student-level results. In this example, each student would take a 10-item test, whether it is administered as a computer-based assessment or in traditional paper-and-pencil mode.

However, at the level of the school, district, or state, the number of forms makes a difference. No matter how they are spread across Forms A through D, the 40 KCA items would provide much better coverage of the "Math Computation" construct than would the 10 P&P items included only on Form A. As a result, the KCA will enable inferences to be drawn or decisions to be made about program or school effectiveness that are much more trustworthy. Being based on only one-quarter as many items, the P&P will provide indicator-level instructional reports that are pretty much worthless. Having only one or two items per indicator would hardly provide a dependable basis on which to group students or adjust instruction.

Thus, having only one P&P form will affect scores and proficiency rates not only in the short run (by precluding the conduct of an appropriate KCA-P&P equating study) but also in the long run.

Minimizing form, mode, and DT effects by equating the KCA and the P&P merely would enable more accurate comparisons of existing proficiency to be made. But, the quality and equity of the indicator-level instructional reports that enable teachers to actually raise the true achievement levels within their schools. In this respect, P&P users will be at a serious disadvantage in the long run. Even if a second P&P form is added next year, as STATE X has promised, the KCA versions of the state assessments would still consist of twice as many items, and the P&P would be half as dependable.

Impact on Precision and Reliability

When combined to make school-level or district-level decision, the four KCA forms will yield results that are more precise and more reliable than would any single form of the test. Greater measurement precision is especially important in two situations: (1) when classifying "borderline" students as either Basic or Proficient and (2) when computing confidence intervals for AYP determinations, when computing Safe Harbor, and when computing confidence intervals for Safe Harbor determinations. Greater measurement precision reduces the risk that a "borderline" student, school, or district will be misclassified.

The standard error of the mean, which represents the spread of individual scores around the average score, is an indicator of measurement precision. The smaller the standard error, the more precise is the measurement. As shown by Exhibit 16, the standard errors from the four forms of KCA will be roughly two times more precise than any one P&P form.

Exhibit 16

Comparison of Standard Errors Derived from One Form and Four Forms (Grade 7 Math – 2004)

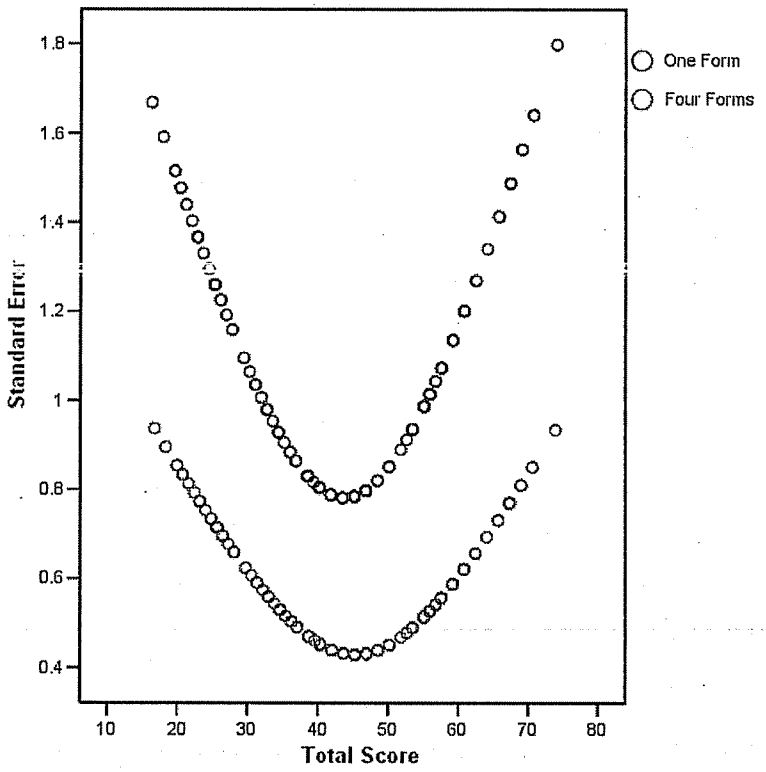
Subject	Grade	Level	Standard Error One Form	Standard Error - Four Forms	Difference	Relative Size (Ratio of 1 Form to 4 Forms)
Math	7	Unsatisfactory	1.31	.70	.57	1.8
		Basic	.89	.51	.39	1.8
		Proficient	.83	.45	.38	1.9
		Advanced	1.09	.58	.52	1.9
		Exemplary	1.50	.85	.72	1.9
		Total	1.07	.59	.49	1.8

Exhibit 16 indicates that standard errors derived from one form will be nearly twice larger than the standard errors derived from combining four forms. Consequently, measurement precision will be half as great. In addition, the observant reader will have noticed the good news: the measurement precision with both one form and four forms is greatest near the cut point between Basic and Proficient. This point is demonstrated visually in Exhibit 17.

Exhibit 17

Scatterplot of Standard Errors Derived from One Form and Four Forms

(Grade 7 Math – 2004)



Although the best level of score precision with one form is not much better than the worst level of precision with four forms, every little bit helps, given that the greatest disparities between the KCA and P&P proficiency rates occur in the score range between the Basic and Advanced levels, especially on either side of the cut point between the Basic and Proficient levels, as illustrated in Exhibit 18.

Exhibit 18

Control Graph of the Difference between the KCA and P&P Proficiency Rates Along the Total Score Continuum (Grade 4 Math – 2004)

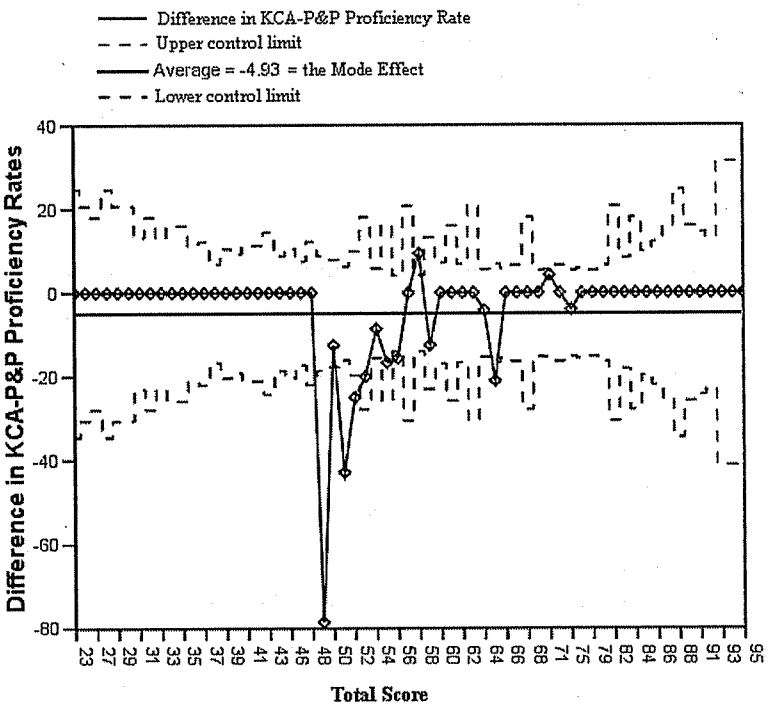


Exhibit 18 is a control chart. Control charts serve as a graphical aid for evaluating the variability in a manufacturing or quality control process, such as the production of electronic components or the management of water pollutant levels. But they are readily applicable to evaluating differences in proficiency rates. By distinguishing between acceptable and unacceptable degrees of variability in the process of assigning students to performance levels, one can identify where the process may need adjustment.

It is worth noting that the red line extending horizontally from – 4.93 on the vertical Y-axis represents the mode effect – the average difference in grade 4 math proficiency rates across the entire score scale. It shows that, on average, the KCA proficiency rate was 4.93 percentage points less than the P&P proficiency rate. Meanwhile, the black line that begins to fluctuate at the Basic-Proficient cut point of 48 on the Total score scale depicts the DT effect.

For easier interpretation, Exhibit 19 replaces the Total score continuum with the five state assessment performance level classifications.

Exhibit 19

Control Graph of the Difference between the KCA and P&P Proficiency Rates Along the Five Performance Level Classifications

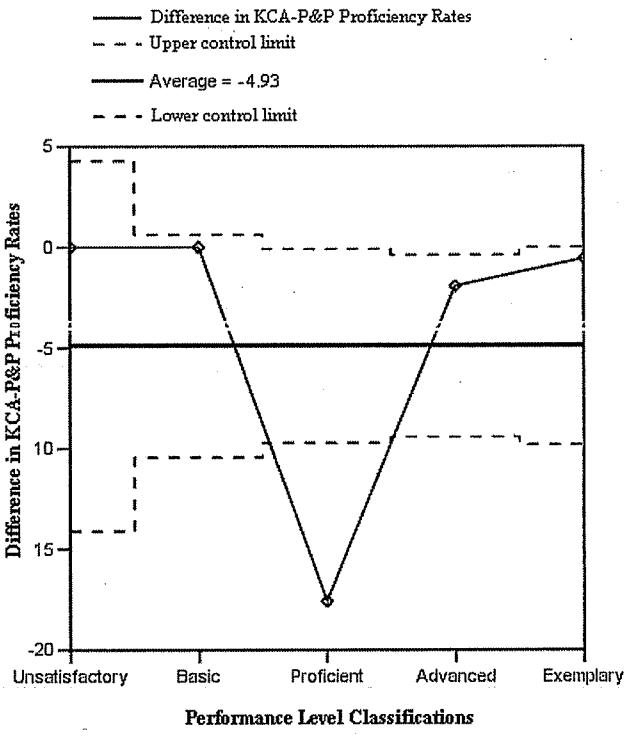
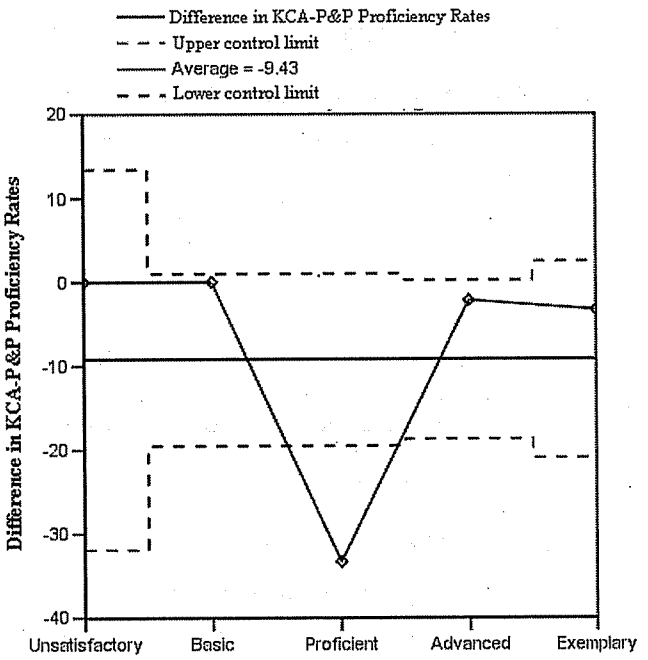


Exhibit 19 clearly shows that the process of classifying students into performance levels is most out of control right at the point where the greatest accuracy would be most desirable – at the delineation between Proficient and Below-proficient.

The additional lack of measurement precision that will ensue from having only one P&P form but four or more KCA forms will only exacerbate the situation, as indicated in Exhibit 20.

Exhibit 20

Control Graph of the Difference between the KCA and P&P Proficiency Rates With Only One P&P Test Form



The pattern in Exhibit 20 is the same as the pattern in Exhibit 19. But the magnitudes of the differences have increased. The Mode effect has increased from 4.93% to 9.43%. The difference at the Proficient-Below Proficient distinction has increased from slightly more than 15% to more than 30%.

The disparity in proficiency rates between the KCA and the P&P has been serious enough for the last two years without doubling it. Yet, a doubling of the disparities between KCA and P&P is what will likely occur with only one P&P form but four or more KCA forms.

Impact on Scores and Proficiency Rates

When combined to make school-level or district-level decisions, four test forms will provide estimates of performance better than will any single test form with respect to not only precision but also form equivalence. This is so because, even after the test forms are equated on a statewide basis, between-form differences still exist at the district and school levels. With four forms, the differences tend to "average out" toward the middle. But with only a single form, no such averaging out can occur. So, if the P&P test-takers are assigned to one of the more difficult forms of the new assessment, they will clearly be at a disadvantage.

Consider the Grade 8 reading scores of the 48 students who double-tested in 2004 at Marshall Middle School. Their results are presented in Exhibit 21.

What if only one form of the P&P had been available – say, Form 80?

Exhibit 21

Grade 8 Reading – 2004: Marshall Middle School

Form	P&P mean Total score	KCA mean Total score
28		76.0
45		83.5
62		80.5
80	79.6	77.1
Average	79.6	79.3

For Form 80, the average P&P score was 79.6, while the average KCA score was 77.1 – a difference of 2.5 points. Given only one P&P form, the average P&P score is also 79.6. However, with four KCA forms, the "averaging out" leads to an overall KCA average score of 79.3 – only 0.3 points lower than the P&P score.

So, how does one form versus four forms affect the P&P and KCA proficiency rates?

The general relationship between the mean score and proficiency rate can be derived from a regression analysis and expressed as:

(Total x Slope) + Intercept = Proficiency

Thus, for XYZ's grade 8 results in 2004:

P&P Proficiency = (79.6 x 5.128) – 363.499 = 44.69 % Proficient
KCA Proficiency = (79.3 x 4.185) – 271.404 = 60.47 % Proficient

Bearing in mind that the AYP goal for grade 8 reading in 2004 was 57.3% proficient, the KCA test-takers as a group made AYP while the group of students who took Form 80 via P&P did not make AYP. The easier KCA forms balanced out the more difficult KCA forms.

So, what would have happened if four P&P forms had been available?

Exhibit 22 provides all of Marshall Middle School's actual double-test scores from 2004.

Exhibit 22

Grade 8 Reading – 2004: Marshall Middle School

Form	P&P mean Total score	KCA mean Total score
28	77.7	76.0
45	84.5	83.5
62	83.6	80.5
80	79.6	77.1
Total	81.3	79.3

Form 28 was more difficult than Form 80. Form 45 and Form 62 were easier. Combined, they yielded an overall average score of 81.3. In turn, the overall average yielded a proficiency rate of 58.3%. This is one percentage point above the AYP target of 57.3%.

P&P Proficiency = (81.3 x 4.278) – 289.478 = 58.32% Proficient

Further, having only one P&P form compared with multiple KCA forms will have an adverse effect on the stability of scores and proficiency rates across years. Having multiple test forms tends to "average out" not only between-form differences in a single year but also fluctuations in performance over time. Consider the grade 5 reading proficiency rates at Lewis Elementary.

Exhibit 23

Lewis Elementary – Grade 5 Reading Proficiency in 2004 and 2005 by Form

	Year	
Form	2004	2005
19	66.7%	61.5%
36	50.0%	78.6%
Overall	58.3%	70.4%

From 2004 to 2005, the proficiency rate at Lewis decreased by 5.2 percentage points on Form 19 but increased 28.6 percentage points on Form 36. With two forms, the decrease and the increase tended to "average out" – that is, they offset each other – yielding an overall proficiency rate that increased by a more modest but still praiseworthy 12.1 percentage points. Given the random distribution of test booklets to students within classrooms, the ability levels of the students who took each form were probably fairly

comparable. Thus, the differences in difficulty are more likely associated with the specific content included on each form of the test. Consequently, if Lewis had administered only Form 19, the school would have experienced a decline in performance, and it would have failed to make AYP. Conversely, if Lewis had administered only Form 36, the school's performance would have appeared overly inflated.

Clearly, having only one P&P form compared with multiple KCA forms places P&P test-takers at a disadvantage. Clearly, the P&P and the KCA are not substantively comparable. Clearly, the state has broken its promise to maintain a dual-mode state assessment that is equitable.

PART III: Conclusions and Recommendations

The Consequences

The state's declaration that the KCA and the P&P are statistically comparable – that they yield interchangeable results – was based on a small study conducted in 2003 that examined only the average grade 7 math scores of regular education students who took the test without accommodations. The findings of the state study led the KCA in 2004 and 2005 to be authorized for full implementation at all grade levels, with all students, and with all test types. The state assessments have thus underestimated the achievement and progress of many students, many schools, and many districts. Many schools undoubtedly did not make AYP or Standard of Excellence because they did not test each of their students in the mode that would have elicited his or her optimal and thus more valid performance.

For two years, the state has encouraged schools to double-test as many students as possible so that data for comparability studies – and, if indicated, equating studies – could be conducted. The data have not been used. Yet, despite a new generation of assessments about to debut, the state has recently decided not to conduct a comparability study in 2006. As a result, the lack of statistical comparability between the KCA and the P&P will persist. School and district proficiency rates will continue to be underestimated.

In addition, the state recently announced that only one P&P form of the new assessment will be available in 2006. This will perpetuate the inequity in the dual-mode system. Decisions based on P&P results will be less valid and dependable than inferences drawn from the KCA because the P&P will provide less content coverage and poorer construct representation. Assessment by the P&P will be less precise and thus less reliable than assessment by the KCA. Because there will be only one form, scores, performance level classifications, and proficiency rates yielded by the P&P will be unable to "average out" across forms or over time. Being based on far fewer items, the P&P will provide instructional planning information that is much less trustworthy than what the KCA will provide.

Combined, the lack of equating and the difference in the number of test forms means that both P&P and KCA schools will continue to look worse in the short run than they actually are. In the long run, P&P schools will be disadvantaged with respect to making instructional improvements that could truly raise student achievement. Because the two modes will be neither statistically nor substantively comparable, districts and schools would be foolish not to convert entirely to KCA.

The only catch is this: many districts will need to invest in new computer hardware and infrastructure to implement full conversion to the KCA mode. The investment would likely require a significant reallocation of resources away from instruction and other current priorities. Smaller districts would need to spend thousands of dollars. For larger districts, the cost could run into six or seven figures. For districts

such as XYZ or ABC or DEF, full conversion to the KCA could cost from $8 to $16 million – more if the conversion were to be implemented in a single year.

And even then, unless the KCA is equated to adjust for both the mode and double-test effects, the state assessments will continue to underestimate student performance levels and thus school and district proficiency rates.

Granted, these conclusions are based on analyses of only XYZ's data. So, examine your own district data (please contact me if you would like assistance) and draw your own conclusions.

Preferably, the state would collect fresh double-test data in 2006 via a rigorous repeated-measures research design with appropriate counterbalancing and tracking of administration sequence. Preferably, the state would then authorize an independent expert to evaluate the comparability of the KCA and the P&P and, if indicated, to derive equating coefficients that would adjust for both the mode effects and the DT effects for application to the results of all schools.

We would nominate Michael Kolen at the University of Iowa to conduct that independent comparability study. Not only is Dr. Kolen one of the nation's leading experts on scaling and equating techniques, but he also is a member of the STATE X Technical Advisory Committee, the group of experts responsible for providing STATE X and its contractor with advice and assistance regarding the technical quality of the STATE X state assessments. Alternatively, we would recommend that The Buros Institute at the University of Nebraska be contacted.

In Conclusion

The STATE X Assessment Advisory Council (KAAC) and the STATE X Technical Advisory Committee (TAC) are hereby urged to recommend that STATE X take the following actions.

(1) Provide the same number of P&P forms as there are KCA forms.

(2) Equate the KCA and the P&P to adjust for both the Mode effect and the DT effect.

Further, KAAC members are hereby encouraged to counsel their district's superintendents to unite so that they can collectively lobby STATE X to address these issues.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999*), Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Psychological Association (1986*). Guidelines for Computer-Based Tests and Interpretations*, Washington, DC: American Psychological Association.

Bennett, R.E. (2003). *Online assessment and the comparability of score meaning.* Princeton, NJ: Educational Testing Service.
See also: http://cresst96.cse.ucla.edu/products/overheads/BennetREVISED.pps

Court, S.C. (2006, April). *The Interchangeability of Dual-Mode Assessment Results*. Paper accepted for presentation at the annual meeting of the American Educational Research Association: San Francisco.

Henry, S.A. & San Martin, T. (2004, September). "STATE X Assessment Advisory Council." Presentation to the STATE X State Test Coordinators' Conference: Salina, KS.
http://www.STATE X.org/assessment/kaachenrysanmartin.ppt

Linn, R.L. and Kiplinger, V (1994). *Linking Statewide Tests to the National Assessment of Educational Progress: Stability of Results CSE Technical Report 375.* CRESST/University of Colorado at Boulder.
http://cresst96.cse.ucla.edu/CRESST/Reports/TECH375.PDF

Paek, P. (2005). *Recent Trends in Comparability Studies.* PEM Research Report 05-05 Iowa City, Iowa: Pearson Educational Measurement.
http://www.pearsonedmeasurement.com/downloads/research/RR_05_05.pdf

Poggio, J., Glasnapp, D.R., Yang, X. & Poggio, A.J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *The Journal of Technology, Learning, and Assessment, 3*(6), 1-30.
http://www.bc.edu/research/intasc/jtla/journal/pdf/v3n6_jtla.pdf .)

Poggio, J, Consolver, R. (November 28, 2005). "Correspondence to STATE X Superintendents and Test Coordinators." Lawrence, KS: Center for Educational Testing and Evaluation (CETE).
http://www.STATE X.org/assessment/assessregistration2006.pdf

Porter, A. C. (1991). "Assessing national goals: some measurement dilemmas." In T. Wardell (Ed.), *The assessment of national goals. Proceedings of the 1990 ETS Invitational Conference.* Princeton, NJ: Educational Testing Service.

Randall, C. (Dec. 5, 2005). E-mail correspondence.

Office of Elementary and Secondary Education, (2004) *Standards and Assessments - Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001.* Washington, D.C. United State Department of Education.
http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf (See especially p. 38.)

Wang, T. & Kolen, M.J. (2001). "Evaluating comparability in computerized adaptive testing: Issues, criteria and an example." *Journal of Educational Measurement, 38*(1), 19-49.