Final
Report
HumRRO
FR-00-26
June 2000
Revised

School Classification Accuracy Final Analysis Plan for the Commonwealth Accountability and Testing System

R. Gene Hoffman
Lauress L. Wise

DRAFT until released by KDE and NTAPAA

Prepared for:
Kentucky Department of Education
Frankfort, KY 40601

Contract Number M-00003669

Human Resources Research Organization
66 Canal Center Plaza, Suite 400 • Alexandria, VA 22314-1591

School Classification Accuracy Final Analysis Plan for the Commonwealth Accountability and Testing System

R. Gene Hoffman and Lauress L. Wise

Human Resources Research Organization

The purpose of this report is to describe two alternative methods for analyzing school classification accuracy for the interim accountability cycle for the Commonwealth Accountability and Testing System (CATS). The report targets a technical audience, in general, and Kentucky's National Technical Advisory Panel on Assessment and Accountability (NTAPAA), in particular. The report assumes that the reader is familiar with CATS, but it does begin with a brief review of some of the most salient design features in order to ensure understanding of the proposed analyses. At the outset, it must be understood that the testing design precludes optimum analyses that take into account all of the factors that might impact on schools' accountability scores. Therefore, both proposed approaches rest on assumptions that may have different degrees of credibility. In addition, both designs focus on scores from the Kentucky Core Content Test even though there are other components in the interim accountability model. The assumptions, limitation, and strengths of both approaches are described. A final decision about which approach to pursue will be made in the June, 2000 NTAPAA meeting.

Salient Features of CATS for the Interim Accountability Cycle

The CATS interim accountability cycle begins with the school year of 1996-1997 and ends with the school year 1999-2000. Because testing occurs in the spring of each school year, we will reference each year with the spring date only. Data for 1997 and 1998 constitute the "baseline" years upon which targets scores will be set for the 1999 and 2000 school years. For the interim cycle only, targets will be set using a regression approach in which the average of 1997 and 1998 accountability indexes will be correlated with the average of 1999 and 2000 accountability indexes. Details are available elsewhere (Carlson, 1999). The import aspect of this interim model is that schools' scores for 1999 and 2000 are being compared to each other, with the comparison being adjusted for performance in the baseline year. For parsimony, regression results will be regarded as fixed in our analysis of measurement error. Regression error (i.e., standard error of *prediction* vs. standard error of *measurement*) has already been incorporated in the interim cycle accountability decision model.[1]

Schools' accountability index for each school year includes scores from each of seven content areas of the Kentucky Core Content Test (including writing) and a score for non-academic indicators. The non-academic index (formerly the non-cognitive score) tends to be much higher than the academic index scores, and also tends to have much less variance across

[1] Schools who perform better than their predicted performance will be rewarded, with the benefit of doubt given to schools as long as their accountability index is greater than their predicted index. Schools who perform below their predicted score will be offered assistance with a safety band of one standard error of prediction. Only schools that are one standard error of prediction below their predicted scores will be offered assistance.

schools. The influence of unreliability from the non-academic index on classification accuracy will not be investigated by either analysis method. Sufficient (student level) data is not available and the anticipated impact is not sufficient to make acquisition of such data worthwhile.

At the school level, scores from six content areas of the Kentucky Core Content Test (writing is excluded) are expressed as non-linear transformations of students' scale scores. Based on their Kentucky Core Content Test scale scores and cutpoints set by previous standard setting research (reference technical reports -- KDE, 1995; KDE, 1997), students are assigned to one of four achievement levels: Novice, Apprentice, Proficient, and Distinguished (NAPD). The lower two levels, Novice and Apprentice, are subdivided into thirds (low, middle, and high), resulting is eight total achievement categories. Based on Kentucky statues, points are awarded to these eight categories in the following array (from low Novice to Distinguished): 0, 13, 26, 40, 60, 80, 100, and 140.[2] Currently for most content areas in most grades, the highest percentage of students are in the Novice and Apprentice categories. For each content area, a schools' content index is the average of its students' NAPD points. Because these classifications and scores that are used by the accountability system to build school scores from student scores, they play the primary role in the classification accuracy analyses.

Forms design and test administration design are two more important consideration for the alternative school classification accuracy analyses plans. For the 1999 and 2000 assessments, there are twelve forms for Arts & Humanities and for Practical Living/Vocational Studies, and six forms for each of the remaining content areas. With anomalous exceptions, items within the forms do not overlap and students take only one form. In other words, within each school, students and forms are completely confounded. The distribution of forms is spiraled with the intention that this will create randomly equivalent groups of students. Scale scores for each form are calibrated/equated with each other on the basis of this assumption. In other words, distributions of scale scores should be approximately equal.[3] The forms, however, are systematically constructed to cover somewhat different aspects of the Core Content for Assessment (KDE, 1999) so that the set of forms for a given content area provides a more representative distribution of items then any one form. The extent to which this differential targeting of content creates student-by-form interactions (i.e., form differences that are greater than would be expected from within-form, student-by-item interactions) is not knowable given the testing design.

Overview and Objective of the Analyses

The objective for each of the analysis approaches is to build standard errors of measurement (SEMs) for schools' interim cycle 1999/2000 target scores. Standard errors of measurement can then be applied to each school's observed index and used to calculate the probability that its true score would place it in a different school classification. The first, and somewhat more complex approach, is based on the previously developed student observed score classification accuracy analysis (Hoffman and Wise, 1999, 2000a, 2000b). The result of

[2] While the array is non-linear with respect to student proficiency, it can be considered linear with respect to the value placed on proficiency, where value has been set by legislation.

[3] There are some caveats to this description. The first and second forms are targeted for the visually and hearing impaired students. These students are not included in calibration runs. CTB has proposed a modification to the forms design to incorporate a small number of overlapping items across forms.

classification accuracy approach will be a series of standard errors of measurement that vary by potentially three factors: size of school, the academic index itself, and within school variance of students. The second approach is a generalizability theory analysis elaborated after Yen (1997) and Miller (1999). This approach will yield one standard error of measurement for each of several selected school sizes.

For either approach, we will analyze each Kentucky Core Content Test grade/subject combination separately deriving standard errors for each separate academic index. These standard errors can then be combined to estimate the standard error for the combined academic index which is used in the overall accountability computation. By separating grade/subject estimates, school-level standard errors can be estimated for any configuration of grades within a school.

Caveat on the reliability of difference scores

For each school, interim cycle accountability decisions have essentially two data points: an average index score for the cycle baseline (1997 and 1998), and an average index score for target years (1999 and 2000). Both scores contain measurement error. Furthermore projecting from the correlations for mid-year interim regression (Carlson, et al, 1999), scores for baseline and target years are expected to be highly correlated. Ironically, the high correlation allows for good prediction of schools' scores for 1999/2000 from their 1997/1998 scores, yet the reward system is based on the extent to which schools vary from their prediction. In a sense, the reward system is like a phantom multiple regression: the first variable in the equation is past performance, and the "phantom" second variable is instructional change. However, the strong observed correlations between past and target performance leave little room (i.e., valid residual variance) for detecting instructional change. For example, Carlson et al's mid-year estimates for middle schools show that the 1997/1998 index average correlates .87 with 1999 index scores. Using the formula for the correction for attenuation due to unreliably and assuming that the reliability for the average 1997/1998 index is equal to the reliability of the 1999 index, then the lowest reliability possible for the index scores is .87. However, when the reliabilities of two scores are each .87, then a formula provided by Cronbach and Furby (1970) shows that the reliability of their difference score is zero. On the other hand, if the reliabilities each happen to be .99, then the reliability of the difference score is .92. There are two importance points. First, variations in reliabilities of difference scores are greatly magnified from small variations in the reliabilities of the component scores. The second and more important point for our analysis is that the *estimation* of reliabilities of differences scores are greatly impacted by over- or under-estimation of the reliabilities of the component scores. Regression residual scores are difference scores.

The language of the statue may provide some relief in the sense that it appears to declare past history as fixed, such that classification accuracy is a function of only the reliability in the target, and not baseline, years.[4] The NTAPAA, however, must wrestle with whether to accept or reject this premise. Our two alternative analysis can each produce standard errors of measurement for the combined target years (1999/2000). Formulas (adapted from Ghiselli,

[4] Personal communication, Scott Trimble, May 22, 2000.

1964) will be presented to forecast standard error for difference scores between baseline and target years.

Some Assumptions that Necessary for Both Approaches

Several assumptions that are required to make the alternative approaches feasible. The assumptions below are shared by both approaches.

First, one of the significant criticisms of the old Kentucky Instructional Results Information System (KIRIS) was that cohorts of students varied sufficiently in their abilities and prior achievements that year-to-year the comparison of school means based on different cohort was invalid (Hoffman, 1998). Both approaches use within-school, within-year calculations of student variance to estimate cohort sampling effects. This is based both on the assumption that each school is drawing from its own population, so that within-school student variance is expected to be less than total student variance, and therefore, provides an adequate estimate of year-to-year cohort variation. It is also based on the reality of the data.

We have no independent assessment of the extent to which cohort-to-cohort variation (1) may be a result from true cohort population differences (a changing neighborhood), (2) are the result of other effects that are independent of school interventions (e.g., differences in cohort achievement norms that may arise from unique leadership patterns within the cohorts), or (3) are simply within the range of variation in group means that can be predicted for variance in individual scores. In other words, by using within-year student variance to estimate cohort variance, we may under- or over-estimate cohort variance or be reasonably accurate. An interesting study, which is outside of the scope of the present work, would be to make an ANOVA-like comparison between year-to-year school (i.e., cohort mean) variance to the cohort mean variance predicted by the variance of students (divided by student sample size).

A second important assumption which will be employed by both approaches is that errors of measurement are uncorrelated across students, across the different grade/content assessments, and across years. In some respects, this is very reasonable. For example, item sampling within one content area, which could impact unreliability due to school-by-forms interactions, would not be expected to correlate with unreliability due to item sampling in another content area. On the other hand, because there is overlap in pupils taking different portions of the assessment, there may be student-level characteristics that systemically affect scores in the multiple assessments that the students have in common. Whether these characteristics are classified as measurement error or as valid correlates of achievement would be open to interpretation. Our analysis assumes that student characteristics that covary with achievement on overlapping assessments are not a source of error. This assumption, obviously, rests on ethnicity and gender bias being adequately controlled during test construction.

A third assumption has already been mentioned. Student-by-forms interactions cannot be studied, so that the extent to which any student-by-forms error variance affects school level scores cannot be directly studied either. On the other hand, at the school level, all schools receive all forms, and forms are (or at least are intended to be) randomly assigned to students. Since forms are be randomly distributed among students, any unknown student-by-form interactions should, in theory, balance at the school level of analysis. Obviously, this

assumption would be expected to hold more tightly for larger schools. At the school level, school-by-forms interaction signals a mix of two elements: content that is intended to be addressed differently by different students (valid variance) and error variance from the sampling of items within the forms. Again, these cannot be separated, so some assumption about whether to call form interactions valid or error is necessary. Yen (1997) compared the observed school-by-form differences to expectations based on test construction, and concluded that the data she was examining did not support the assumption that school-by-forms interactions could be accepted as valid variance. She therefore treated it as error variance. That will be our approach as well.

Approach A: School-Level Standard Errors of Measurement based on Student Classification Accuracy Results

The first analytic approach builds from our previous student observed classification accuracy method and results (Hoffman & Wise, 2000a, 2000b). This method estimates probabilities of unknown true achievement being in any of the four different NAPD proficiency categories, given observed scale scores. These probabilities can be used to calculate probabilities that schools' index scores could actually be higher or lower than their observed index scores, if students' true classifications were actually known. The student classification accuracy effort, however, analyses only the major NAPD categories and does not include the Novice and Apprentice subdivisions. This adds a simplifying assumption to this approach, in addition to the assumptions described above. The analyses will begin with separate analyses for each grade/subject combination.

School level analyses by grade and subject

The primary engine for these analyses will be an iterative process for each school which synthesizes a distribution of possible school-level academic index scores based on our student-level accuracy statistics. Based on his/her observed score, each student has some probability of contributing points to his/her schools which match each of the four NAPD categories. Repeating for each student within a school, potential total points will be accumulated for the school in proportion to the odds of the student contributing those various point values. After all students within a school are processed, and the results adjusted for number of students in the school, the result in a probability distribution of potential school-level true scores based on student-level expected true scores. For each school, the mean ($Mn_{exst/cg}$) and variance ($\sigma_{exsc/cg}^2$) of this potential school-level true-score distribution will be calculated (see Table 1 for notation and further explanation). $\sigma_{exsc/cg}^2$ is an estimate of the effect of student measurement error on school-level scores.

In addition, the student accuracy data will be used to calculate on expected true-score classification for each student. For each school, the variance for these student-level expected true scores will then be calculated ($\sigma_{exst/cg}^2$). $\sigma_{exst/cg}^2$ divided by school sample size will be used as an estimate of school-level measurement error due to student sampling. More homogeneous schools will be projected to have less student sampling error.

Finally, we know from the student accuracy demonstration (Hoffman & Wise, 2000a) that student-level measurement has a "regression to the mean" bias. That is, expected true scores

are neither as high nor as low as observed scores. Therefore, schools with large concentrations of students on one end of the achievement continuum or the other may have true scores there are less extreme than their observed scores. This systematic bias can be capture by the difference between a school observed index ($Mn_{obst/cg}$) and their expected true score ($Mn_{exst/cg}$).

Table 1
Statistics for school classification accuracy compiled for student classification accuracy

Abbreviations: ob = observed score
ex = expected score
st = student
sc = school
c = content area
g = grade

Designation	Description
$Mn_{obst/cg}$	School mean of the students' observed scores = content area academic index for a school
$Mn_{exst/cg}$	$Mn_{etst/cg}$ = *expected true* school index based on projected distributions of possible students' true scores. Note that the unusual terminology ("expected true") is not redundant. From our set of probabilities of students having true score value in different performance level categories, we will build probabilities of various *possible* school true index values. The mean of the possible true index values is the *expected true* index, given students' assigned scores.
$Mn_{obst/cg}$ - $Mn_{exst/cg}$	Expected bias in school observed index. Note: Student scoring is constructed such that the *expected* true score for an observed scale score is systematically biased. Expected true scores are more near the center of the distribution then assigned scale scores, i.e., there is a regression toward the mean effect. Therefore, schools with large numbers of low scoring students will have their true scores underestimated, and schools with large numbers of high scoring students will have their true scores overestimated.
$\sigma_{exsc/cg}^2$	Within school variance of possible school-level true indexes. $\sigma_{etsc/cg}^2$ is centered on school's expected true score, not observed school index. See note on $Mn_{exst/cg.}$ $\sigma_{etst/cg}^2$ is the potential variance in a school scores given our uncertainty about students' true scores, i.e., variance in school score due to student measurement error.
$\sigma_{exst/cg}^2$	Within school variance of students' expected true scores.
$\sigma_{exst/cg}^2 / n_{p:s}$	Sampling error for school level scores
$\sigma_{scerror/cg}^2$	Distribution of expected school true index, given student observed scores = adjustment for bias + measurement error + sampling error $= (Mn_{obst/cg} - Mn_{etst/cg})^2 + \sigma_{etsc/cg}^2 + \sigma_{etst/cg}^2/N$

Thus, the three error components will be calculated:

- Expected variance in school's accountability indexes due to student sampling.
- Expected variance in school's accountability indexes due to student measurement error.
- Expected difference between school's observed index and expected true index due to systematic bias in students' observed scores.

Each of these components is calculated for each school. These three error components can be combined as indicated in Table 1 to estimate composite error variance for each school for each of its applicable grade and subject indexes, $\sigma_{scerror/cg}^2$.

Modeling of error variance across all schools

Although it is feasible to report unique error estimates for each school, it may be too complicated and, therefore, not credible to do so. For this analytic method, we also propose to model the relationship between the school-level composite error estimates ($\sigma_{scerror/cg}^2$) and three particular school characteristics: school size, mean observed student achievement level, and within school variance in students' observed achievement. School size may be expected to relate to error because of its relationship to the sampling component of our error composite. Level of student achievement is included in the modeling because of its potential relationship to the bias component of our error composite. Finally, within-school student variance is include because is related to the student sampling component of our error composite. We will plot our school-level error composite against each of the three school characteristics, and attempt to fix polynomial regression equations.

Some degree of interpretation will be needed after the data is available and analyzed to determine how to best present a set of school error estimates that adequately represent what we observe in the relationships between our three school characteristics and the school-level error estimate. A three-way classification of schools with multiple level category (e.g., large, medium, and small size schools crossed the high, medium, and low school scores, crossed with high and low student variability) starts to become more complicated than simply reporting unique error estimates for each school. We will have to depend on the data being informative of which school characteristics make enough difference in estimating school-level measurement error to matter.

Combining error estimates across years

The above analyses will be conducted for each of the 1999 and 2000 school years. For each grade and subject, 1999 and 2000 standard error estimates will need to be combined to give a single standard error estimate. We have made the assumption that measurement errors are independent, therefore adding the separate 1999 and 2000 error estimates is appropriate. Since the target year score is an average of 1999 and 2000, the error variance for the composite is the sum of the separate error variance components divided by number of years squared, which is 4 in this case. The results at this point in the analysis would be 18 separate standard errors of measurement, one of each of the grade/subject combinations, each of which represents the expected error in academic index scores for the 1999 and 2000 years combined.

Approach B: School-Level Standard Errors of Measurement based on Generalizability Theory Analysis

The alternative to estimating standard errors from student classification accuracy data is to conduct generalizability analyses of student NAPD scores. Given that school scores span two years, the basic model is one in which pupils are nested within form and within schools and form are crossed with school.

Figure 1 presents the design and Table 2 and 3 presents the calculations using Brennan's (1981) notation and algorithms for generating sums of squares and variance components. For each of the grade/subject combinations, the six sources of variance in schools' two-year academic index averages include: (1) school, (2) year, (3) school by year, (4) form within year, (5) school by form within year, and (6) pupil within form within school by form. The order of the nesting terms in the last source of variance is a little ambiguous in its wording since pupils are nested within forms, within schools, and within years. However, for derivation of the error components, the expressed order of the nested does not matter, as long as the nesting is captured.

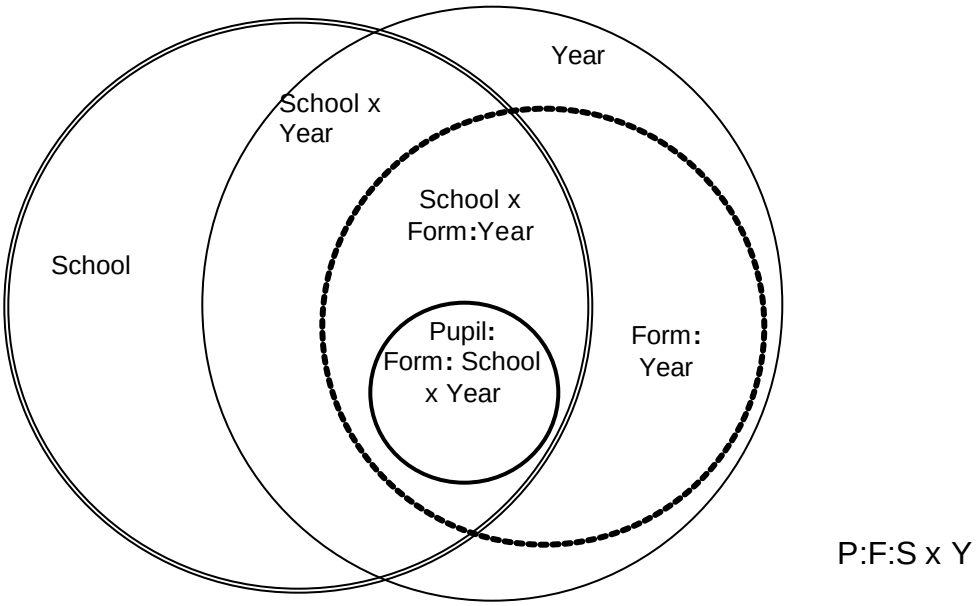

Figure 1. Generalizability theory design representing Kentucky Core Content Test two-year accountability cycle.

Table 2
Estimating Variance Components for Pupil: School Year Form Generalizability Theory Design – Random Effects Estimates

Effect	df	Means	SS
School (s)	$n_s - 1$	$\bar{X}_s = \frac{1}{n_y n_f n_p} \sum_y \sum_f \sum_p X_{syfp}$	$n_f n_y n_p \Sigma \bar{X}_s^2 - n_s n_y n_f n_p \bar{X}^2$
Year (y)	$n_y - 1$	$\bar{X}_y = \frac{1}{n_s n_f n_p} \sum_s \sum_f \sum_p X_{syfp}$	$n_s n_f n_p \Sigma \bar{X}_y^2 - n_s n_y n_f n_p \bar{X}^2$
School x Year	$(n_s - 1)(n_y - 1)$	$\bar{X}_{sy} = \frac{1}{n_f n_p} \sum_f \sum_p X_{syfp}$	$n_f n_p \Sigma\Sigma \bar{X}_{sy}^2 - n_f n_y n_p \Sigma \bar{X}_s^2 - n_s n_f n_p \Sigma \bar{X}_y^2 + n_s n_y n_f n_p \bar{X}^2$
Form:Year (f:y)	$n_y(n_f - 1)$	$\bar{X}_{f:y} = \frac{1}{n_s n_p} \sum_s \sum_p X_{syfp}$	$n_s n_p \Sigma\Sigma \bar{X}_{yf}^2 - n_s n_f n_p \Sigma \bar{X}_y^2$
School x Form : Year (sf:y)	$n_y(n_s - 1)(n_f - 1)$	$\bar{X}_{sf:y} = \frac{1}{n_p} \sum_p X_{syfp}$	$n_p \Sigma\Sigma\Sigma \bar{X}_{syf}^2 - n_f n_p \Sigma\Sigma \bar{X}_{sy}^2 - n_s n_p \Sigma\Sigma \bar{X}_{yf}^2 + n_s n_f n_p \bar{X}_y^2$
Pupil: School Year Form (p:sfy)	$n_y n_s n_f (n_p - 1)$	na	$\Sigma\Sigma\Sigma\Sigma X_{psyf}^2 - n_p \Sigma\Sigma\Sigma \bar{X}_{syf}^2$
Total	$n_s n_y n_f n_p - 1$	$\bar{X} = \frac{1}{n_s n_y n_f n_p} \sum_s \sum_y \sum_f \sum_p X_{syfp}$	$\Sigma\Sigma\Sigma\Sigma X_{psyf}^2 - n_s n_y n_f n_p \bar{X}^2$

Table 3
Estimating Variance Components for Pupil: School Year Form Generalizability Theory Design – G-Study Estimates

Effect (α)	Estimated σ^2 –Random Effects Model	Estimated $\sigma^2(\alpha\|M)$ -- Mixed Models (N = Universe size)	
		Basic Mixed Model	Year Fixed
School (s)	$\frac{[MS(s) - MS(sy)]}{n_y n_f n_p}$	$\hat{\sigma}^2_s + \frac{\hat{\sigma}^2_{sy}}{N_y} + \frac{\hat{\sigma}^2_{sf:y}}{N_f N_y} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_f N_y N_p}$	$\hat{\sigma}^2_s + \frac{\hat{\sigma}^2_{sy}}{N_y}$
Year (y)	$\frac{[MS(y)-MS(sy)-MS(fy)+MS(sfy)]}{n_s n_f n_p}$	$\hat{\sigma}^2_y + \frac{\hat{\sigma}^2_{sy}}{N_s} + \frac{\hat{\sigma}^2_{f:y}}{N_f} + \frac{\hat{\sigma}^2_{sf:y}}{N_s N_f} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_s N_f N_p}$	$\hat{\sigma}^2_y$
School x Year	$\frac{[MS(sy) - MS(sfy)]}{n_f n_p}$	$\hat{\sigma}^2_{sy} + \frac{\hat{\sigma}^2_{sf:y}}{N_f} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_f N_p}$	$\hat{\sigma}^2_{sy}$
Form:Year (f:y)	$\frac{[MS(fy) - MS(sfy)]}{n_s n_p}$	$\hat{\sigma}^2_{f:y} + \frac{\hat{\sigma}^2_{sf:y}}{N_s} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_s N_p}$	$\hat{\sigma}^2_{f:y}$
School x Form : Year (sf:y)	$\frac{[MS(sfy) - MS(syfp)]}{n_p}$	$\hat{\sigma}^2_{f:sy} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_p}$	$\hat{\sigma}^2_{f:sy}$
Pupil: School Year Form (p:sfy)	MS(syfp)	$\hat{\sigma}^2_{p:f:sy}$	$\hat{\sigma}^2_{p:f:sy}$

Table 4
Estimating Variance Components for Pupil: School Year Form Generalizability Theory Design – D-study Estimates

Effect (α)	D-study error component	Use term in	
		Absolute error estimate	Relative error estimate
School (s)	$\hat{\sigma}^2_s + \frac{\hat{\sigma}^2_{sy}}{N_y}$		
Year (y)	$[\hat{\sigma}^2_y / N_y]\ [1 - \frac{n_y}{N_y}] = \mathbf{0}$	(X)	
School x Year	$[\hat{\sigma}^2_{sy} / N_y] \times [1 - \frac{n_y}{N_y}] = \mathbf{0}$	(X)	(X)
Form:Year (f:y)	$\hat{\sigma}^2_{f:y} / N_y N_f$	X	
School x Form : Year (sf:y)	$\hat{\sigma}^2_{f:sy} / N_y N_f$	X	X
Pupil: School Year Form (p:sfy)	$\hat{\sigma}^2_{p:f:sy} / N_y N_f n_p$	X	X

Random, fixed, or sampled from a finite universe

Generalizability theory explicitly considers the universe to which observed score are interpretable. Typically, the items that make up a particular test are only viewed as samples of an infinite array of similar items. Being sampled from an infinite domain, items are therefore considered "random." On the other hand, some facets may cover the intended universe to which scores are intended to generalize. Year, for example, could be considered fixed because the universe of generalization is two years and both years are sampled. On the other hand, year could be considered as sampled from a finite universe. The logic is this: The school academic index, while directly interpretable as the average of students' achievement, is being used to make inferences about the instructional programs of those schools. An accountability cycle is four years long. Changes in instruction that occur in any of those four years could impact students' achievement in the final two years. Thus, the universe of generalization could be viewed as instructional change that occurred in any of the four years of the cycle. Only two of the four years are assessed, however. Other than being illustrative of sampling within a fixed domain, we are making no strong argument that the present data be treated with years being samples of a fixed four-year domain. Instead, we are suggesting that school and years be considered fixed. Forms and pupils are assumed to be randomly sampled from a infinite domain. Table 4 indicates that the value of for two sources of variance (year and school x year) reduce to zero when years are considered fixed.

Note that current literature is mixed on whether pupils should be considered fixed, random, or sampled from a fixed domain (Cronbach, Linn, Brennan, & Haertel, 1997; Hambleton, Jaeger, Koretz, Linn, Millman, & Phillips, 1996; Yen, 1997). Persistent criticisms of Kentucky's accountability model that cohort-to-cohort variation in student proficiency is unfair (Hoffman, 1998) makes treating students as fixed unwise. Yen uses two different approaches, one for which students are random, and a second for which students are treated as samples of a finite domain with that domain being defined as the total school population from which the tested students are taken. Yen's second approach does not fit Kentucky's two year cycle very well, particularly since we know the transience among students is perceived to be a significant issue for some districts (Thacker, Koger, Hoffman, and Koger, 2000) and is indeed related to school scores (Medsker, 1998). Therefore, we have chosen to treat students as random, i.e., sampled from an infinite universe. (Note also that in Yen's second approach, she adds a term for measurement error at the person level. That term is mathematically eliminated when students are treated as random.)

Yen (1997) also discussed potential modification to the forms by schools interaction given that forms are intended to target slightly different content. She concludes that since there is no way to directly test differences in targets (forms and students are confounded), the straightforward approach, as presented in Tables 2 – 4, is more acceptable with a caveat that it may overestimate standard error.

Absolute and relative error

Generalizability theory considers two kinds of error: absolute and relative. Absolute error is appropriate to consider when the objects of measure (schools in our case) are being assessed against a standard that generalizes beyond any of the particular instances of the various

facets of measurement (e.g., different forms, different years, different pupils). Relative error, on the other hand, is appropriate when schools are being compared to each other and have been subject to the same measurement processes (same forms, same years). Table 4 indicates which variance component enter each type of error estimate. With years treated as fixed, three error components (form within year, school by form within year, and pupil within form within school by form) are summed to estimate absolute error. Only the later two components (school by form within year, and pupil within form within school by form) are summed to estimate error variance for the relative model. Because the regression model does compare schools to each other, the relative error model is more appropriate for interim cycle accountability scores.

Conduct separate analyses for selected school sizes

Because the number of pupils within each school varies tremendously and has a significant impact on estimation of variance components, we will conduct the analyses with three representative categories of schools based on class sizes. Representative sizes will be selected separately for the $4^{th}/5^{th}$ grade combination, the $7^{th}/8^{th}$ grade combination, and the $10^{th}/11^{th}$ combination. Representative sizes will be selected by dividing schools into thirds based on class size, and identifying the median size for each of those third, that is, the $1/6^{th}$, 1/2, and $5/6^{th}$ points in the size distribution. The results of the analyses to this point will be three error variance estimates, one for each selected school size, for each of the 18 grade/subject combinations.

A Note on Rater Error

The open response items for Mathematics, Reading, Science, Social Studies, Arts & Humanities, and Practical Living/Vocational Studies are all hand-scores by raters. No direct assessment of rating errors are included in the analyses for these subject, however. The effects of rater errors are confounded with item sampling errors and so are captured as part of forms effects and school-by-form interactions, and as part of the true score projections from the student classification accuracy results. No further attempt will be made to isolate the effects of rater errors for these subjects.

Writing

Up until this point in the discussion, we have ignored students' writing scores. They represent a seventh content area and are composed of two separate scores, one for portfolios and one of the on-demand writing prompt. Both of these are scored holistically by as few as one rater per student. Both of these assessment are also averaged into schools accountability scores, and therefore can contribute both reliable and unreliable variance. Because of the way they are scored, neither assessment can be analyzed by form the student classification accuracy approach. Both can be addressed, however, by some form of generalizability analysis.

Each student completes one on-demand writing prompt, and it is chosen by the student from a pair of alternatives. Six pairs of writing prompts constitute six forms for on-demand writing. From past analysis (Hoffman, Koger, & Awbrey,1997), we know that means for different writing prompts vary greatly for prompts within a form as well as for prompts from different forms. The variation in means leads to the conclusion that each prompt should be treated as a separate "form" using the same generalizability analysis design described above. As

far as the self-selection factor in concerned, we see no option other than considering it one of the random factors affecting prompt (i.e., item) sampling.

Portfolios, however, are (in theory[5]) unique to each individual student. "Forms" as a theoretical facet for portfolios is confounded with students.[6] Therefore, school-level error variance for portfolios will be assessed using a generalizability design similar to the one presented above, but without form as a facet. That is, pupils are nested within the intersection of schools and years. Formulas for this three facet (pupils:schools x years) are available in Brennan (1981), designated as i:(p x h) in his notation.

Combine Grade/Subject Standard Errors of Measurement

Whether we derived error variance estimates from the student classification results or from generalizability theory, to this points we will have SEMs (square roots of the error variances) only at the grade/subject level. These need to be combined to estimate the error in total school accountability scores which is a weighted combination of academic index scores, as appropriate for the school, from selected grades and subjects. That is, for typical elementary schools, each of six separate content tests, plus the two forms of writing, are administered at either the 4th or 5th grade level. Therefore, the separate error variance estimate must be combined to form a single estimate of error for the total academic index. Kentucky statute specifies the weights for each subject, and these weights vary somewhat for different grade levels. Nevertheless, because we are assuming that errors of measurement are independent, we can additively combine our error variance estimates, using weights that reflect a combination of the statute weights and the numbers of components included in the composite accountability index.

At this point in the analysis, we will have reached our initial objective of having standard errors of measurement for total accountability index scores for the target years. Assuming that errors are normally distributed, the SEMs can be converted to probabilities that estimate the likelihood that a schools' true classification is within the assigned classification, is in a lower classification, or is in a high classification, if we assume that target scores are fixed without error.

Extension to Difference Score Standard Errors

Should the NTAPAA reject the supposition that target scores are fixed, then some further analysis will be required to estimate standard errors for difference scores.

Student level classification accuracy results are currently available only for 1999, and student classification results for 2000 will have to be calculated prior to the school level analysis. In order to estimate school classification accuracy of difference scores based on student classification results, we will have to make some assumptions about the 1997 and 1998 data.

---

[5] Some schools do tend to structure common activities and present selected topics for students to create portfolio entries.

[6] Again, this is an oversimplication. Anecdotally, some schools reportedly been doing a better job than others of structuring portfolio activities that facilitate higher quality writing. "Item sampling," therefore, may be confounded with schools. In this unusual case, schools become both the object of measurement and an instrument, or facet, of measurement.

Our working assumption is that because the 1997 and 1998 assessments did not operationally score the multiple choice items, individually these years may be less reliable than either 1999 or 2000, but the combination of the two years should be at least as reliable as either 1999 or 2000. Therefore, we would assume for this analysis that the reliability for the 1997/1998 average equals the reliability of 1997 alone. In addition, prior to double-weighting the open-response items, there were an equal number of score points for multiple-choice and open-response items in 1999 and 2000. Therefore, in 1997 and 1998, there were one-half as many score points in the assessment. Under an assumption that multiple-choice and open-ended score points are equal in their contribution to measurement reliability, then the Spearman-Brown prophecy formula can be used to step down the two-year 1999/2000 reliability to an estimate for the 1998/1997 baseline index.

The generalizability perspective is also somewhat different because test design changed in 1999. During the 1997 and 1998 testing years, the forms contained common items. However, since students were scored with all items in the form they took, there is a form effect. A simple approach would be to conduct analyses on 1997 and 1998 test scores exactly parallel to generalizability analyses described for the 1999 and 2000 test data.

The interim regression model essential rewards schools based on their residual scores, which is simply a difference score which equals:

(1999 index+2000 index) index – a ((1997 index+1998 index)/2) + b,

where a and b are the regression slope and intercept.

It follows that variance for the residual is

$$\sigma_{sc(average1999/2000)}^2 + a^2 \cdot \sigma_{sc(average1997/98)}^2 - 2 \cdot a \cdot \sigma_{sc(average1999/2000)} \cdot \sigma_{sc(average1997/98)} \cdot r_{average/average},$$

and the error variance for the residual is $\sigma_{scer(average1999/2000)}^2 + a^2 \cdot \sigma_{scer(average1997/98)}^2$, where σ_{sc} is total variance for the designated year(s) and σ_{scer} is error variance for the designated years.

References

Brennan, R. L. (1981). *Elements of Generalizability Theory.* Iowa City, IA: ACT Publications.

Carlson, J. E. (November 10, 1999). *Kentucky Accountability Regression 1997-98 to 1999 Prediction.* Monterey, CA: CTB/McGraw-Hill.

Cronbach, L. J. & Furby, L. How should we measure "change" – or should we? *Psychological Bulletin, 74,* 68-80.

Cronbach, L. J. Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57,* 373-399.

Ghiselli, E. E. (1964). *Theory of Psychological Measurememt.* New York: McGraw-Hill.

Hambleton, R. K., Jeager, R. M., Koretz, D., Linn, R. L., Millman, J., & Phillips, S. E. (1995). *Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991-1994.* Frankfort, KY: Office of Educational Accountability, Kentucky General Assembly.

Hoffman, R. G., Koger, L., & Awbrey, A. (1997). *Changes in spelling, capitalization, punctuation, and subject/verb agreement skills under the Kentucky Education Reform Act* (LRS97-1). Frankfort, KY: Bureau of Learning Results Services, Kentucky Department of Education.

Hoffman, R. G., & Wise, L. L. (1999). *Establishing the Reliability of Student Level Classifications: Analytic Plan and Demonstration.*(FR-WATSD-99-34). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G., & Wise, L. L. (2000a). *Establishing the reliability of student proficiency* classifications*: The accuracy of observed classifications.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, April, 2000. Also available at www.Humrro.org.

Hoffman, R. G., & Wise, L. L. (2000b). *The Accuracy of Students' Novice, Apprentice, Proficient, and Distinguished Classifications for the Kentucky Core Content Test.* Alexandria, VA: Human Resources Research Organization.

Kentucky Department of Education, 1995. *KIRIS Accountability Cycle 1 Technical Manual.* Frankfort, KY: Author.

Kentucky Department of Education, 1997. *KIRIS Accountability Cycle 2 Technical Manual.* Frankfort, KY: Author.

Kentucky Department of Education, 1999. Core Content for Assessment. Frankfort, KY: Author. Available at http://www.kde.state.ky.us/oapd/curric/corecontent/core_content_index_version_30.asp

Medsker, G. J. (1998). *Determining the Relationship between Student Transience and KIRIS School Results: Are Schools with Transient Students Unfairly Impacted?* (HumRRO Report FR-WATSD-98-12). Radcliff, KY: Human Resources Research Organization.

Miller, M. David. (April, 1999). *Generalizability of Performance-Based Assessments at the School Level.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, April, 1999.

Thacker, A. A., Koger, L. E., Hoffman, R. G., & Koger, M. E. (2000). *The Transition from KIRIS to CATS, Year 2: Instruction, Communication, and Perceptions at 31 Kentucky Schools* (FR-WATSD-99-23). Alexandria, VA: Human Resources Research Organization.

Yen, W. M. (1997). The technical quality of performance assessments: Standard Errors of Percents of Pupils Reaching Standards. *Educational Measurement: Issues and Practice, 16,* 5-15.