DRAFT Final Report: An Investigation of ELDA Performance Level Scores under Conjunctive and Compensatory Models

August 26, 2010

Produced under contract with the CCSSO by Metrica Research Associates, LLC. This report includes material encompassed by contract line items identified as specifications, interim reports 1 and 2, and final report. Metrica is grateful to Professor Edward Haertel of Stanford University for his contribution to the design of the methodology for this report. The author (William Lorié of *Metrica*) is responsible for all recommendations and any errors, however.

Background and Purpose

In Summer 2010, the ELDA SCASS (English Language Development Assessment State Collaborative on Assessment and Student Standards) of the Council of Chief State School Officers (CCSSO), in consultation with its technical advisory committee (TAC), commissioned a study to explore the impact and desirability of switching the scoring model of the newly-produced ELDA short forms (ELDA-s). The current model is conjunctive, and the model being explored is compensatory.

A prior study commissioned by the ELDA SCASS (ELDA: Four Studies, 2009) argued for the desirability of implementing a compensatory model as part of ELDA revision, the other part of the revision being a reduction in test length for three of the four components tests (Speaking, Listening, Reading and Writing). The SCASS proceeded with the form reduction process (ELDA: Speaking Study Report, 2010; ELDA: Short Form Selection Report, 2010) accordingly, but reserved judgment on the means by which cut scores would be implemented on ELDA-s.

The SCASS commissioned the present study to better understand the implications, on its new short forms, of continuing with its current conjunctive model or switching to a compensatory model.

The study consisted of two parts. The first was the derivation of an appropriate compensatory model. The second part was an assessment of classification accuracy for the two models under consideration, for both the long and short forms. Form A of ELDA-long (ELDA-l) was chosen to conduct the study. Because ELDA forms are parallel, results should generalize to other forms. On key statistics, such as cut points on the logit (or *theta*) scale derived from item parameter estimation, ELDA forms are treated as equivalent.

The study specifications called for an investigation encompassing all three grade spans of ELDA (grades 3 to 5, grades 6 to 8, and grades 9-12); however, the unavailability of verifiable item response theory (IRT) parameters for the Writing items for the grade 9-12, Form A, made it impossible to assess accuracy at grade 9-12 under the study methodology. The author does not believe this undermines the conclusions of study, in large part because of the similarity of results between the first two grade spans.[1]

Methodology

ELDA is a complex testing program in the sense that it is comprised of several elements and procedures for score determination. For the central (grades 3 through 12) component of ELDA, there are separate tests for each of the grade spans 3-5, 5-8, and 9-12. For each of these tests, there are four components subtests: Speaking, Listening, Reading, and Writing. Each of these subtests within grade-span has a psychometric scale, and since there are three parallel forms per grade-span, score transformation tables exists for each form within subtest within grade-span.

During a standard setting study in 2005 (ELDA: Standard Setting Study Report, 2006), cut scores were set on each subtest for the one operational form at the time (Form A), resulting in five performance levels, ranging from 1 to 5, per subtest. Although only one form was used to set performance levels, the cut scores were expressed for each subtest on the underlying theta metric, which is independent of test form.

ELDA sponsors also decided at the time of standard setting on a conjunctive model for determining a composite score, also ranging from 1 to 5. The conjunctive model treated the combination of performance levels on each subtest in such a way that performance on some of the subtests was more critical to high scores on the composite.

For reasons outlined earlier, in 2009 ELDA sponsors began considering the desirability of a compensatory model for determination of the composite. In general, compensatory models are more appropriate when the components under consideration are strongly related theoretically or empirically.

Data Preparation

Data from the last administration of Form A was requested of all seven current participating ELDA states. Because of miscommunications, one state could not send data in time for the study and the data for another state needed to be dropped

[1] Nevertheless, the unavailability of verifiable items parameters for Writing, Grade 9-12, Form A, is troubling because it prevents certain forms of equating typical for IRT testing programs, for that form.

because it could not be reconciled with the required format. Data for the remaining states were formatted in a standard format for file merger and cases were dropped where any one of the component raw scores was missing. The resulting number of cases is displayed in Table 1.

Table 1. Student Data Case Counts

	Grade 3-5	Grade 6-8	Grade 9-12
Students	23,531	15,483	12,177

Derivation of a Compensatory Model

If a conjunctive model maps four *component* PLs to a *composite* PL, then so should an alternative compensatory model. One approach to deriving a compensatory model would be to treat each component PL as continuous and regress the composite PL onto the vector of component PLs. The result would be a prediction equation that fits the composite PL as a weighted linear function of component PLs.

This approach is sensible because it estimates the relative importance of each component PL to the composite, but it can be biased unless each of the 625 cells in the 4-dimensional space of component PLs is appropriately weighted. The S-L-R-W pattern 5-5-5-5 is less common than, for example, 3-4-3-3, with each number representing PLs for (S)peaking, (L)istening, (R)eading, and (W)riting.

The weighting of cases in each of the 625 cells was determined by the real data. Thus, the regressions to derive the compensatory model were run on the cases in Table 1, separately for each grade-span. To convert regression coefficients into weights summing to 1, the regressions were run so as to suppress the constant terms, and the resulting coefficients were each divided by the sum of coefficients. This procedure establishes a continuous *intermediate composite performance level* (ICPL) scale for each grade-span, with values ranging from 1 to 5.

Mapping values of ICPL onto a final composite PL is equivalent to turning the ELDA conjunctive model into a compensatory model. This was done by determining "cut points" on ICPL for each of levels 2, 3, 4, and 5.

It was considered desirable by ELDA sponsors to determine cut points on a compensatory model of ELDA in such a way that proficiency distributions on the conjunctive model could be preserved to the extent possible.

Under the assumptions that (1) the cost of misclassification is the same regardless of threshold and (2) the cost of misclassification is the same whether a student is classified into a higher level (false positive) or lower level (false negative) under the compensatory model (and relative to the conjunctive model) then minimization of

the sum of false positives and false negatives as a function of ICPL would yield the appropriate ICPL cuts for each threshold.

This procedure was applied to each threshold (2, 3, 4, 5) for each ICPL function (one each for span 3-5, 6-8, and 9-12) to derive the compensatory cuts.

Evaluating Classification Accuracy

Classification accuracy measures the extent to which observed classifications differ from true classifications. The difference between observed and true classifications is based on a measurement model where measurement error in observed scores, especially for students close to a category threshold, can cause observed scores to differ from true scores in such a way that the category assigned according to the observed score would be different than the category that would be assigned had one known the student's true score.

All tests with less than perfect reliability have less than perfect accuracy. Since one does not have access to true scores, the classification accuracy of ELDA was assessed by simulation. The simulation procedure entailed generating a group of 50,000 "plausible examinees" using the following information by drawing pseudorandom 4-element vectors from a distribution with the variance-covariance structure of the real ELDA data. Each element of the vector represents a true theta score for the S, L, R, and W subtests. The variance-covariance matrix of the thetas, $\boldsymbol{\Sigma}$, was used because it is known that the scales are correlated. And, although technically the variance-covariance matrix appropriate for this procedure is the *true* $\boldsymbol{\Sigma}$, the *observed* variance-covariance matrix was used instead. This is because estimation of the true $\boldsymbol{\Sigma}$ is more complicated, relies on further assumptions, and would likely not have yielded results with different practical implications.

A set of plausible examinees was generated for grade spans 3-5 and 6-8 separately. (Again, it was not useful to carry this procedure out for 9-12 because of problems with the existing Writing item parameters, which would be needed in later steps.)

Since the thetas thus generated could be treated as true, and cut scores are ultimately expressed on the theta metric (see ELDA: Standard Setting Report, 2006), a vector of true component PLs could be generated for the plausible examinees. In addition, because item parameters and the IRT scoring model were known (Rasch for the dichotomous items and Masters Partial Credit Model for the polytomous items), it was possible to generate item response vectors for each subtest, for each the 50,000 plausible examinees per grade span, sum the scores for each subtest, and apply existing raw-to-PL tables to get manifest (or observed) component PLs. Sometimes the true PLs were the same as the manifest PLs, and sometimes they were different.

At this point, either scoring rule (conjunctive or compensatory) could be applied to yield the *composite* PL for each plausible examinee. Both were applied an evaluated separately in terms of classification accuracy.

Finally, a variation was introduced to understand the classification accuracy for the short forms, ELDA-s. This consisted in systematically removing from the plausible examinee response vectors each item deleted during the short form selection process, applying the appropriate short form raw-to-PL table, and repeating the analysis. This step yielded the classification accuracy of the short forms.

At the end of the classification accuracy phase of the study, results for accuracy were obtained for ELDA-l conjunctive, ELDA-l compensatory, ELDA-s conjunctive, and ELDA-s compensatory.

Evaluating Classification Consistency

Classification *consistency* is a measure of the degree to which two measurement procedures, each resulting in classifications of examinees, lead to similar classifications. It can be argued that classification consistency is of greater practical importance than accuracy in assessing the desirability of switching scoring models. If the new model predicts a meaningful difference in the distribution of students across performance levels, then the validity of the switch may be called into question. (That said, however, consistency is meaningless without accuracy.)

To evaluate classification consistency, classification consistency tables were generated for the plausible examinees of the accuracy study. Consistency was measured on the *observed* composite PLs under both scoring rules in order to generalize to available real data.

A global measure of consistency was derived by summing the percent of students classified into Level *x* for both the conjunctive rule and the compensatory rule, for each level of *x* (that is, 1, 2, 3, 4, and 5).

As a check, the procedure for classification consistency was carried out for one of the real ELDA tests, for both the long and short forms.

Results

Classification Accuracy

This study found that the classification accuracies of ELDA Composite performance levels (CPLs) were very high for the current conjunctive model, and that reduction of the forms did not diminish that accuracy to any great extent. (Classification accuracy is defined as the percent of students who are both truly at level *x* and are actually classified into level *x* by the test instrument and its scoring rule, summing

across all performance levels *x*). See Table 2 for classification accuracy results for the conjunctive rule current used by ELDA.

Table 2. Classification Accuracy Results, Simulated Data, Conjunctive Rule

	Grade 3-5	**Grade 6-8**
ELDA-l	81.7%	82.9%
ELDA-s	80.2%	82.1%

This study found that, if (1) weights determined by a regression of observed composite PL on the observed component performance levels are used to produce a continuous *intermediate composite performance level* (ICPL), based on a weighted sum of the component performance levels; and (2) cut points are set on the ICPL metric in such a way as to minimize classification inconsistency between the conjunctive-rule CPLs and the compensatory-rule CPLs, then a high percentage of students would be accurately classified by the compensatory rule. In other words, a compensatory rule determined by observed performance levels is at least as accurate as the conjunctive rule. See Table 3. This is important because the procedure proposed for determining the new compensatory performance levels on the real data was the same one used on the simulated data. Accuracy drops negligibly for the short forms.

Table 3. Classification Accuracy Results - Simulated Data, Compensatory Rule

	Grade 3-5	**Grade 6-8**
ELDA-l	82.2%	82.9%
ELDA-s	80.6%	82.1%

Classification accuracy results support a decision to adopt ELDA-s regardless of whether a conjunctive or compensatory model is used.

Classification Consistency

Classification consistency is defined as the percent of students who would be classified into the same level under two different procedures. In the case ELDA, the two different procedures correspond to (1) the existing conjunctive rule and (2) a compensatory rule determined as described above.

This study found that, if (1) weights determined by a regression of composite PL on the component performance levels are used to produce a continuous *intermediate composite performance level* (ICPL), based on a weighted sum of the component performance levels; and (2) cut points are set to be equal to the value of ICPL that minimizes classification inconsistency between the conjunctive-rule CPLs and the compensatory-rule CPLs, then a high percentage of students would be classified *consistently* by the two rules. (Note: This is similar to a similar-sounding result reported above, but this time it is about consistency, not accuracy.) The high levels

of consistency are in great part due to the fact that ELDA's conjunctive rule approaches a composite rule in many respects.

The classification consistency evaluation method was implemented on long and short forms of simulated data to verify its feasibility for ELDA. Results are in Table 4.

Table 4. Classification Consistency Results - Conjunctive and Compensatory

	Grade 3-5	Grade 6-8
ELDA-l	94.6%	94.8%
ELDA-s	94.3%	94.5%

Because the method can be applied directly on observed PL distributions, results were verified for long and short forms A for one of the data sets (Grades 3-5), with PL weights and ICPL cut points computed from the real data. The result was a 95.1% level of consistency for both the ELDA-l and ELDA-s.

PL weights and ICPL cut points differ by grade-span but follow a general trend. The weights are shown in **Error! Reference source not found.**. All regressions determining these weights exhibited very high levels of fit, with adjusted R-squares exceeding 0.98 and all variables statistically significant at the 0.001 level or better.

Table 5. ICPL Weights

	Grade 3-5	Grade 6-8	Grades 9-12
Speaking	0.03205	0.07435	0.04151
Listening	0.10466	0.09910	0.07175
Reading	0.43359	0.40512	0.46584
Writing	0.42970	0.42144	0.42090

For a compensatory rule to match the current ELDA conjunctive rule, Reading and Writing should carry the greatest relative weight. Whether weights are determined using simulated or real data, weights for Speaking and Listening are the lowest. This reflects the fact, captured by the regression analyses, that Reading and Writing scores contribute most to a student's composite performance level under the current conjunctive model.

As described in the methodology section, ICPL thresholds, or in other words cut points on the ICPL scale, were determined separately by grade span through a procedure that maximizes classification consistency. For the real data, here the level thresholds are display in Table 6. As a practical matter in moving from a conjunctive to a compensatory system, the weights in Table 5 would be applied to the PLs for S, L, R, and W, respectively, the results added to obtain a continuous-valued ICPL between 1 and 5, and the resulting ICPL classified into a final CPL based on the thresholds in Table 6.

Table 6. ICPL Thresholds

	Level 2	Level 3	Level 4	Level 5
Grade 3-5	1.75	2.74	3.71	4.86
Grade 6-8	1.69	2.86	3.84	4.90
Grade 9-12	1.70	2.71	3.74	4.93

As always, the component PLs are determined through the appropriate raw-to-PL conversion. For ELDA-l, these are the raw-to-PL tables that ELDA is currently using.