Scores and Scales:
Considerations for PARCC Assessments

Michael J. Kolen
The University of Iowa

June 13, 2012

This paper focuses on considerations involved in choosing scores and scales for the PARCC assessments (PARCC, 2010). The issues considered include the choice of raw score, item response theory (IRT) scores, transformations to scale scores, the number of scale score points, and practical issues in choosing scores and scales. This paper focuses on developing scales within grades rather than on vertical scales. Kolen (2011) considers issues associated with developing vertical scales for the PARCC assessments.

This paper follows from a presentation to the PARCC Technical Advisory Committee on May 3, 2012. Robert Brennan made a companion presentation at this meeting that focused on different score scale decisions in the context of a numerical example. Following the presentation, he developed a paper (Brennan, 2012) that is a companion to the present paper.

Characteristics of the PARCC assessments create psychometric challenges to the development of within grade score scales. The PARCC assessments are mixed-format tests that contain both selected-response and constructed-response tasks. In addition, it is anticipated that these assessments will contain a performance-based assessment component that is administered following approximately 75% of the instruction for the school year along with an end-of-year component that is given near the end of the school year. Scores over these components likely will be combined to develop a total score. In addition, scores on the PARCC assessments are intended to be used to measure growth for students across a wide range of proficiency and with English language learners. The measurement of student growth at a wide range of score levels requires the precise measurement of students all along the score continuum.

Raw Scores

The *raw score* on a test is a function of the scores on the tasks (Kolen, 2006). In this section, three different types of raw scores are considered briefly. See Kolen (2006) for an extended general discussion of raw scores and Kolen, Tong, and Brennan (2011) for a further technical discussion.

The *summed score* is a simple raw score in which the total score is a sum of the task scores. The use of summed scores considers one score point on one type of task as

equal to one score point on another type of task. Summed scores are used in many current state assessments (Rogers 2011a,b).

For *weighted summed scores*, scores on tasks are weighted differently. For example, task weights might be developed as in done with the Advanced Placement (AP) examinations (Kolen & Hendrickson, 2012), so that the proportion of points associated with each task type is the desired proportion of total points. Weights might also be chosen to maximize test score reliability.

For *pattern scores,* scores depend on the particular pattern of task responses and can be more complex than weighted summed scores. Pattern scores are often used with IRT models such as the three-parameter logistic and generalized partial credit models. Pattern scores are used in some current state assessments (Rogers 2011a,b).

Any of these types of raw scores might be used with the PARCC assessments. Because summed scores are easy to explain to test users and are simple to calculate, if feasible, their use with PARCC assessments would likely facilitate score interpretation by test users. However, summed scores require treating a score point on each type of task as equal to a score point on another type of task, which could overly constrain the test development and scoring process. So, it might be necessary to consider weighted summed scores. Pattern scoring can lead to greater score reliability than summed or weighted summed scoring. The choice of raw score for the PARCC assessments necessarily involves balancing computational simplicity, simplicity in developing tasks and scoring procedures, and score reliability.

IRT Scores

IRT models will likely be used with the PARCC assessments. When using IRT models, it is necessary to choose an IRT score. The use of IRT scores is summarized here. See Kolen, Tong, and Brennan (2011) for a more detailed description of these scores.

Maximum likelihood estimators (MLE) *for response patterns* make use of the entire pattern of task scores. These MLE scores are unbiased (for long tests), and they maximize the precision among unbiased estimators of proficiency. MLE scores do not depend on the distribution of proficiency in the population.

Bayes response pattern scores are shrinkage estimators, meaning the scores are less variable than the proficiency parameters. These Bayes scores are biased, but typically contain less error in estimating proficiency compared to the MLE scores. The Bayes scores for individuals depend on the distribution of proficiency in the population. For example, the Bayes score could differ if an examinee was compared to their gender group rather than to the overall population. One of the most often used Bayes scores in IRT is the Bayes expected a posteriori (EAP) estimator.

Test characteristic function (TCF) *scores* relate summed scores (or weighted summed scores) to proficiency through the test characteristic function. The TCF is unbiased for long tests but with typically more estimation error than the MLE. *Bayes EAP summed scores* (sEAP) (or weighted summed scores) are biased and typically less precise than the Bayes response pattern scores. Under the Rasch model, (a) the MLE and TCF scores are the same and (b) the Bayes EAP and Bayes sEAP scores are the same.

Because of the shrinkage property of the Bayes estimators, the variability of the Bayes scores in the population is less than the variability of MLE scores. In particular,

$$\mathrm{var}(\hat{\theta}_{TCF}) \geq \mathrm{var}(\hat{\theta}_{MLE}) \geq \mathrm{var}(\theta) \geq \mathrm{var}(\hat{\theta}_{EAP}) \geq \mathrm{var}(\hat{\theta}_{sEAP}),$$

where each of the different scores is indicated by $\hat{\theta}$ with a subscript, and θ without a subscript is the proficiency parameter. This relationship is illustrated in the following table taken from Kolen and Tong (2010). As can be seen in this table, the standard deviations of the scores are ordered as in the preceding equation. In addition, the proportions of examinees at each of four proficiency levels are given in the table. As can be seen, the proportions of examinees at the two extreme levels (and especially at Level IV) are influenced by the proficiency estimator used.

Table 1. Effects of IRT Proficiency Estimator on Percent in Proficiency Level

	Proficiency Estimator			
	$\hat{\theta}_{TCF}$	$\hat{\theta}_{MLE}$	$\hat{\theta}_{EAP}$	$\hat{\theta}_{sEAP}$
Mean	–.003	.012	–.002	.000
SD	1.164	1.143	.949	.933
Percent in Proficiency Level				
Level I	22.02	20.77	19.27	19.43
Level II	33.94	35.95	38.70	36.53
Level III	33.53	32.53	33.86	37.20
Level IV	10.51	10.76	8.17	6.84

Note. From Kolen and Tong (2010).

IRT Scoring: Unidimensional or Multidimensional

As described earlier, it is anticipated that the PARCC assessments will contain performance-based assessments and end-of-year assessments that will be administered at different times during the school year. In addition, various task types are used. PARCC could decide to treat all of the tasks as if they assess a unidimensional proficiency. In this case, a unidimensional IRT model could be used. As an alternative, separate unidimensional models could be used with each task type. Then a total score would be calculated as a composite of unidimensional scores. Based on methodology described by Kolen, Wang, and Lee (2012), Kolen and Lee (2011) illustrated the effects on estimates of measurement precision when using a unidimensional model when a test is likely multidimensional

The use of a multidimensional model would be more complex. In addition, the use of a multidimensional model might make equating of scores on alternate forms difficult if PARCC decides to not use any common tasks with the performance-based assessments.

Scale Score Transformations

Raw scores or IRT scores typically are transformed to scale scores to facilitate test use. In addition, the use of a scale, in conjunction with equating procedures, allows for scores from different alternate forms of an assessment to be reported on the same scale. Transformations can be linear or curvilinear. Typically, the range of scores is specified and the scale scores rounded (often to integers). Some of the common transformations are described in this section. Brennan (2012) provided a detailed numerical example of different score scale transformations in the PARCC context.

With linear transformations, scores can be transformed to have a particular mean and standard deviation for a population of examinees. Linear transformations also can be created by fixing two score points, for example – the mean and a cut score – at specified values.

For summed scores, measurement error variability typically is relatively high for the middle scores and relatively low at the extreme scores. For IRT scores, the opposite pattern typically is observed; i.e., measurement error variability is relatively high at the extremes and relatively low for the middle scores. Linear transformations of scores maintain these patterns of measurement error variability.

Nonlinear transformations of scores can be used to change the pattern of measurement error variability. For example, when scores are normalized (nonlinearly transformed to have an approximately normal distribution), the resulting scores tend to have measurement error variability patterns similar to those for IRT scores – i.e., relatively high error variability at the extremes and relatively low measurement error variability for the middle scores. In addition, an arcsine transformation of summed scores can lead to measurement error variability

that is nearly constant along the score scale (see Brennan, 1989, for the use of the arcsine transformation with the operational score scale for the ACT Assessment). Brennan (2012) provides examples of the effects of score scale transformations on the pattern of measurement error variability.

The following figure from Kolen and Lee (2011) provides another example. In this figure, note that the conditional standard error (CSEM) pattern is concave up for the IRT and normal scales, concave down for the linear scale (a linear transformation for the summed scores), and nearly equal along the score scale for the arcsine transformed scores. This example clearly indicates that the pattern of measurement precision is highly dependent on the score scale transformation used.

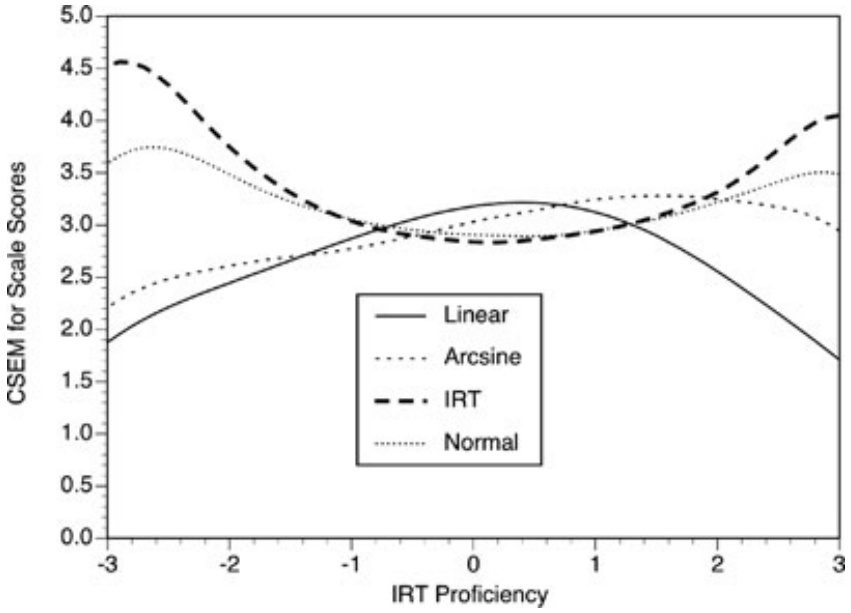

FIGURE 4. Conditional standard errors of measurement for scale scores.

Note: From Kolen and Lee (2011)

Because the PARCC assessments are intended to assess students at a wide range of scores, it might be preferable to stabilize measurement error variability using the arcsine or a similar transformation. Stable error variability might facilitate the assessment of growth and the use of growth models such as value-added models.

Number of Scale Score Points

Kolen (2006) reviewed rules of thumb for deciding on the number of distinct score points to use on a scale. Too few score points can lead to a loss of precision whereas too many score points can lead test users to over-interpret small score differences. Two rules of thumb exist in the literature for choosing numbers of score points. If a test has a reliability of around .9, it can be shown that with 60 score points, a ±3 scale score interval will contain examinees' true scores approximately 68% of the

time. As Kolen (2006) reported, this property led to the choice of 60 distinct score points for the SAT scale. As an alternate rule of thumb, if a test has a reliability of around .9, it can be shown that with 30 distinct score points, a ±1 scale score interval will contain the examinees' true scores approximately 50% of the time. As Kolen (2006) reported, this property led to the choice of 30 distinct score points for the Iowa Tests of Educational Development and 36 score points for the ACT Assessment. These sorts of rules of thumb could be useful for choosing the number of score points to be used with the PARCC assessments.

Choosing Scores and Scales for the PARCC Assessments

Many combinations of scores and scales are currently in use with state testing programs (Rogers, 2011a,b) and in other testing programs, including those examples considered in this paper. The choice of scores and scales can facilitate some test uses and make other test uses more difficult. The goal in choosing scores and scales should be to facilitate the most important uses for a test and to discourage inappropriate uses. The variety of scores and scales used in operational testing programs reflects how test developers valued the importance of various uses.

One trade-off that PARCC will need to consider is that of simplicity versus score precision. If simplicity is emphasized, then summed scores might be preferable with arcsine-transformed scores, and a relatively small number of distinct score points (e.g., 60 score points). However, if the simple approach leads to scores that are deemed not sufficiently precise, then it might be desirable, for example, to use weighted summed scores or pattern scores. Although this paper does not provide a prescription for the type of score scale to use with PARCC, it is intended to provide a framework within which to make decisions about the score scale.

References

Brennan, R. L. (Ed.). (1989). *Methodology used in scaling the ACT Assessment and P-ACT+*. Iowa City, Iowa: ACT Publications.

Brennan, R. L. (2012). *A white paper on scaling PARCC Assessments: Some considerations and a synthetic data example*.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155-186). Westport, CT: American Council on Education and Praeger.

Kolen, M. J., & Hendrickson, A. B. (2012). Scaling, norming, and equating. In C. Secolsky & D. B. Denison (Eds.). *Handbook on measurement, assessment and evaluation in higher education* (pp. 161-177). New York, NY: Routledge.

Kolen, M. J., & Lee, W. (2011). Psychometric properties of raw and scale scores on mixed-format tests. *Educational Measurement: Issues and Practices, 30*(2), 15-24.

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice, 29*(3), 8-14.

Kolen, M. J., Tong, Y., & Brennan, R. L. (2011). Scoring and scaling educational tests. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 43-58). New York: Springer.

Kolen, M. J., Wang, T., & Lee, W. (2012). Conditional standard errors of measurement for composite scores using IRT. *International Journal of Testing, 12*(1), 1-20.

Rogers. H. J. (2011a, November). *Summary of scaling procedures used in statewide assessments*. Unpublished manuscript prepared for USNY Regents Research Fund. Obtained from Scott Marion.

Rogers. H. J. (2011b, November). *Summary of scaling procedures used in statewide high school graduation assessments*. Unpublished manuscript prepared for USNY Regents Research Fund. Obtained from Scott Marion.

Kolen, M. J. (2011). *Issues associated with vertical scales for PARCC Assessments*. Retrieved from http://www.parcconline.org/sites/parcc/files/PARCCVertScale(9-12-2011).pdf.

The Partnership for Assessment of Readiness for College and Careers (PARCC) (2010). *Application for the Race to the Top Comprehensive Assessment Systems Competition*. Retrieved from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf.