A White Paper on
Scaling PARCC Assessments:
Some Considerations and a Synthetic Data Example

Robert L. Brennan
CASMA
University of Iowa

June 10, 2012

On May 3, 2012, the author made a PowerPoint presentation to the Technical Advisory Committee (TAC) for the Partnership for Assessment of Readiness for College and Careers (PARCC) on the subject of "Scaling PARCC Assessments: Some Considerations and a Synthetic Data Example." The PowerPoint slides for that presentation are at the end of this white paper. The purpose of this white paper is to provide context and explanation for the slides.

It is particularly important to note the following caveats.

- This paper does not constitute a set of recommendation about how the scaling of the PARCC summative assessments should be performed. Rather, this paper discusses some important issues that the author believes merit consideration as PARCC considers scaling issues.
- This paper does not address all scaling issues that might be considered for the PARCC summative assessments. For example, in the same session that the appended slides were presented, Kolen provided a separate PowerPoint presentation on "Scores and Scales: Considerations for PARCC Assessments," which treated a number of other issues.
- This paper considers an assessment consisting of multiple-choice questions (MCQ), only, using a set of synthetic data. (Obviously, real data for PARCC are not yet available).
- This paper does not provide many theoretical or computational details; rather the intent is to indicate consequences of particular choices that affect scale-score characteristics.[1]

As noted in Slide 2, the primary focus of this paper is on conditional standard errors of measurement (*CSEM*s). Here, the framework for considering *CSEM*s is closely related to that discussed by Lord (1980, see especially chaps. 5 and 6) in

[1] All computations were performed using a C program written by the author.

the context of item response theory (IRT). Alternatively, a strong true-score model could be employed as discussed, for example, by Brennan (1989), Brennan and Kolen (2004), and Kolen (2006). Issues related to these two approaches are also considered in other references listed at the end of this paper.

Although Lord (1980) often places more emphasis on information functions (*I*) than *CSEM* functions, the basic ideas are closely related. In particular, both *I* and *CSEM* depend on a person parameter and an estimate of it, and results generally depend upon both the *choice* of a person parameter metric and the *choice* of an estimator. There is no psychometrically right answer for either choice, but as illustrated in this white paper, these two choices affect the characteristics of the resulting score scale, and often the affect is rather dramatic.

Item-level Data

The synthetic data for this illustrative example consist of 3PL item parameter estimates for 90 items[2] generated with the following specifications[3] :

- *a* parameters: lognormal distribution with a mean all of -.1393 and a standard deviation of .225,
- *b* parameters: uniform distribution with a minimum of -2.33 and a maximum of 2.33, and
- *c* parameters: beta distribution where A= 4.34; B= 17.66.

Slide 3 provides distributions of the item parameter estimates, as well as a scatterplot of the *a* vs *b* estimates. Slide 4 provides a scatterplot of the traditional point biserials vs. difficulty levels. Note that the traditional difficulty level for all but one item is above .2.

The Parameter θ and two Estimators: Maximum Likelihood and Number-Right Score

As noted by Lord (1980, p. 67), the information function for any score *y* relative to θ is defined as

2 At the time this paper was written, it was anticipated that the summative assessment for mathematics would involve 90 points and consist of an end-of-year (EOY) component that is machine scorable (56 points, some of which are MCQs) as well as a performance-based assessment (PBA) component (36 points).

3 These specifications and resulting item parameter estimates were provided to me on March 13, 2012 by Enis Dogan of Achieve.

$$I\{\theta, y\} = \frac{\left(\frac{d}{d\theta}\mu_{y|\theta}\right)^2}{Var(y|\theta)}, \quad \textbf{(1)}$$

where the numerator is the square of the derivative of μ_y given θ (knowledge of derivatives is not required in this paper).

The denominator of the information function in Equation 1 is the square of the *CSEM*; therefore, the *CSEM* for any score *y* relative to θ is defined as

$$CSEM\{\theta, y\} = \sqrt{Var(y|\theta)}, \quad \textbf{(2)}$$

which is the principal focus of this paper. It is important to note that θ is not restricted to any particular distributional form. Typically, however, it is assumed that for a base form, the distribution of θ is normal with a mean of zero and a standard deviation of one, i.e., *N*(0,1), as depicted in Slide 6.

Assuming θ is *N*(0,1), and treating the simulated item parameters as actual parameters, Slide 7 provides the information functions for two estimates: the maximum likelihood estimate (*MLE*), $\hat{\theta}$, and the number-right score *X* (i.e., the two estimates are $y = \hat{\theta}$ and $y = X$). Note that $I(\theta, X)$ is always less than $I(\theta, \hat{\theta})$, which must be true since the maximum value of the information function is $I(\theta, \hat{\theta})$.

For $\hat{\theta}$ and *X,* Slide 8 provides the *CSEM* functions. Two facts are evident:

- $CSEM(\theta, \hat{\theta})$ is less than $CSEM(\theta, X)$ especially at the lower end of the curve; and
- both curves are concave up ("hold water") which means that *CSEM*s are *larger* in the tails of the distribution of θ for both estimators.

Expected Number-Right Score as the Parameter
Number-Right Score *X* as the Estimator

There is "no unique virtue in the θ scale (Lord, 1980, p. 85)" discussed above. Indeed, any monotonic transformation of θ would serve equally well, as far as the mathematical machinery of IRT is concerned. Probably the most frequently used transformation is the test characteristic curve (*TCC*) that transforms θ to the

expected number-right score (*ENR*) score, sometime denoted ξ. *ENR* is much like (although not quite the same as) true score in classical test theory. For an *n*-item test, the transformation is

$$\xi = \xi(\theta) = \sum_{i=1}^{n} P_i(\theta), \quad \textbf{(3)}$$

where, for the 3PL model,

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}. \quad \textbf{(4)}$$

$P_i(\theta)$ (over all θ) is the characteristic curve for item *i*; so, the *TCC* is the sum of the item characteristic curves.

For the synthetic data in this white paper, Slide 10 gives the *TCC*. Note that *ENR* is a monotonically increasing function of θ; i.e., larger values of θ are associated with larger values of *ENR*. The highest value for *ENR* is 90, the number of items in our hypothetical test; the lowest value is approximately 20. The fact that there are no *ENR* scores lower than about 20 is consistent with the fact that the sum of the *c*'s is about 20 (see Slide 3); note also that there are no very difficult items (recall Slide 4).

Based on the *TCC* in slide 10 and the assumed *N*(0,1) distribution for θ, Slide 11 provides the expected number-right distribution, which is sometimes called the *true* score distribution or the distribution for ξ. Slide 12 provides the corresponding expected *observed* number-right distribution. For both slides the word "expected" means that these are the true and observed score distributions, respectively, that one would expect to get if the data fit the 3PL model perfectly.

The *TCC* in Slide 10 may appear to be a rather gentle transformation, but that perception is deceiving, principally because θ ranges from minus infinity to plus infinity,[4] whereas ξ must be in the range 0 to *n*. That is, to get ξ the θ scores are compressed at the extremes.

One consequence of the severity of the TCC transformation is illustrated by the shape of the $CSEM(\xi, X)$ function in Slide 13. Specifically, the shape is concave

[4] In the slides, for practical purposes, θ is bounded by -4 and +4.

down ("spills water"), which is dramatically different from the concave up shape for $CSEM(\theta, X)$ in Slide 8. Consequently, the choice of metric for the parameter (θ vs. ξ) makes a dramatic difference in the shape of the *CSEM* function.

Slide 8 implies that *CSEM*s tend to be *larger* in the tails of the distribution that uses θ as the person parameter. Slide 13 implies that *CSEM*s tend to be *smaller* in the tails of the distribution that uses ξ as the person parameter. This not a contradiction; it is simply a consequence of choice of metric.

There are situations in which neither one of these choices seems optimal from practical points of view. In particular, it may be desirable that the CSEM function be approximately uniform throughout the score scale range (see, for example Brennan, 1989). This may be desirable for PARCC, since there appears to be considerable concern about measurement precision throughout the score scale range.

Stabilizing Error Variance through use of the Arcsine Transformation

Freeman and Tukey (1950) suggested using the arcsine transformation to stabilize the variance of binomially distributed variables. Subsequently Kolen (1988) suggested using this transformation to stabilize *CSEM*s. Kolen and Brennan (2004, sect. 9.4.3) discuss this issue in considerable detail. Here, the process is discussed and illustrated as two steps:

1. get arcsine-transformed number-correct raw scores, and
2. linearly transform[5] the arcsine transformed raw scores in Step 1 according to some prespecified goals or targets.

Admittedly, Step 2 is a bit vague, at this point. Two examples considered below should clarify matters.

Slide 15 provides a plot of the arcsine transformation for each of the possible number-correct raw scores for our synthetic data example. Letting *sc* stand for score scale, the subsequent slides consider two possible score-scale goals or targets:

1. $sc \pm 3$ covers the true scale score 68% of the time for all *sc* (approximately), which is a goal originally proposed by Truman Kelley and is currently used with the SAT (more or less); and

[5]Linear transformations are chosen here to simplify matters. Non-linear transformations are possible.

2. $sc \pm 1$ covers the true scale score 50% of the time for all *sc* (approximately), which is a goal originally proposed by E. F. Lindquist and is currently used by the ACT Assessment (more or less).

Using $sc \pm 3$ implies that the intended constant *CSEM* is 3, and using $sc \pm 1$ implies that the intended constant *CSEM* is 1. These goals are not quite sufficient, however, to specify a linear transformation. For that we must also specify one point. Here, the arcsine of the middle raw score (90/2 = 45) will be transformed to a scale score of 150, which is abbreviated as "arcsine(45) = 150" in the slides.

The $sc \pm 3$ Case

Consider the $sc \pm 3$ case.

- Slide 17 provides the conversion of raw scores to unrounded scale scores. Note that the conversion is slightly non-linear, as it must be since the arcsine transformation is non-linear.
- Slide 18 provides the relative frequencies for the unrounded scale scores. Note that the effective range is about 66 scale score points, since there is very little frequency below 130 or above 195.
- Slide 19 provides *CSEM*s for unrounded scale scores. There is some variability in the *CSEM*s, but for practical purposes they are nearly constant, and certainly much more so than the *CSEM*s in Slides 8 and 13.
- The *CSEM*s for unrounded scale scores are informative, but they are not likely to be entirely reflective of reported scale scores, which are almost always *rounded*. Slide 20 provides the conversion table for raw to rounded scale scores. Note that there are some two-to-one conversions in the middle of the conversion table and a few gaps at the high end.
- The two-to-one conversions and gaps in the conversion for rounded scale scales cause the *CSEM* function to be somewhat bumpy as indicated in Slide 21. Still, the *CSEM*s themselves are relatively stable at about 3.

The $sc \pm 1$ Case

Consider the $sc \pm 1$ case. Slide 23 provides the raw-to scale score conversion for unrounded scale scores, and Slide 24 provides the conversion for the rounded scale scores. The range of scale scores is about 32, but for unrounded scale scores there are a number of many-to-one conversions throughout the score scale range.

Slide 25 provides the rounded and unrounded *CSEM* functions. As is to be expected, the rounded scale scores are more bumpy than for the $sc \pm 3$ case, but the *CSEM*s are still reasonable close to 1.

Relative Frequencies for Rounded Scale Scores

Slide 27 provides relative frequencies for expected observed scores for the $sc \pm 3$ case. Each of the distinctly higher frequencies in the middle regions of the distribution are caused by two raw scores converting to a single scale score (recall Slide 20). These types of distributions are seldom reported in the literature or technical manuals, but they occur relatively frequently in practice.

Slide 28 provides relative frequencies for the $sc \pm 1$ case. The distribution does not look quite as unusual as the distribution for the $sc \pm 3$ in Slide 27, primarily because for the $sc \pm 1$ case almost all raw scores are involved in a many-to-one conversion (see Slide 24).

Concluding Comments

Standard errors of measurement are clearly scale dependent---a fact that is widely known. It is less well known, however, that *CSEM*s can have vastly different functional forms depending on the person parameter and estimator chosen. It follows that the magnitude and functional form of *CSEM*s are partly under the control of the investigator or policy maker who chooses the person parameter and estimator. This only one of many examples of a basic fact: psychometrics is not solely the prerogative of psychometricians!

References

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

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and square root. Annals of Mathematical Statistics, 21, 607-611.

Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice, 7*, 29-36.

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., & Brennan, R. L. (2004*). Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

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.

Slide 1

Scaling PARCC Assessments: Some Considerations and a Synthetic Data Illustration

Robert L. Brennan
CASMA
University of Iowa

May 3, 2012

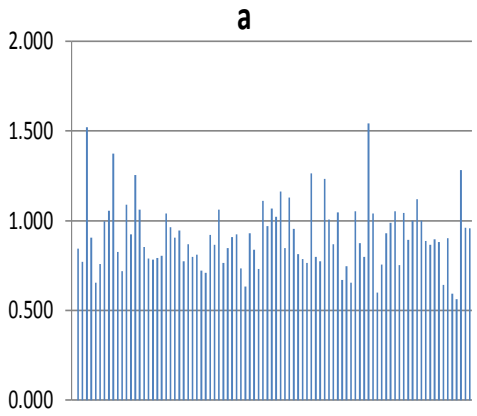
Slide 2

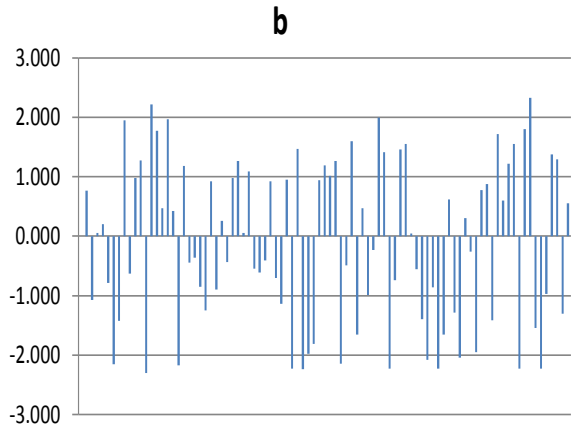
Basic Ideas

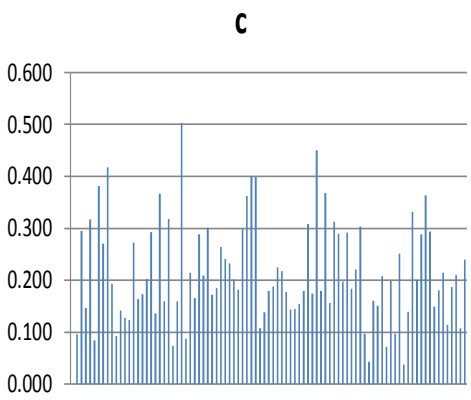
- Information (Person Parameter, Estimator)
  - E.G. I(Θ,y) as discussed extensively by Lord (1980) *Applications of Item Response Theory*
- CSEM (Person Parameter, Estimator) where CSEM means "Conditional Standard Error of Measurement"
- Different results depending on BOTH choice of person-parameter metric AND choice of estimator.
- Psychometrics cannot specify "right" choices
- Choices directly affect characteristics of score scale
- Characteristics for reported score scale are crucial
- See Kolen and Brennan (2004) *Test Equating, Scaling, and Linking* (especially pp. 336-354)

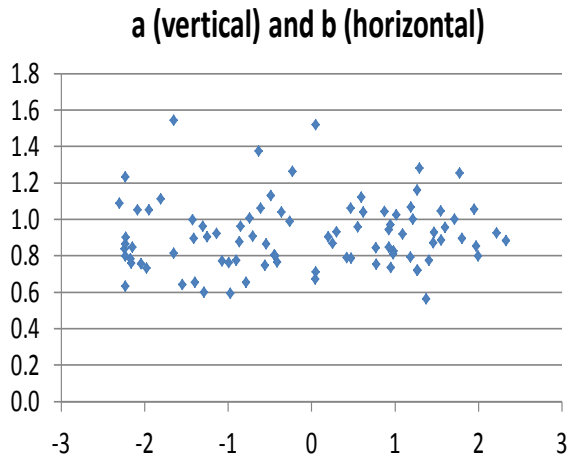
Slide 3

IRT Item Parameter Estimates (90 Items)



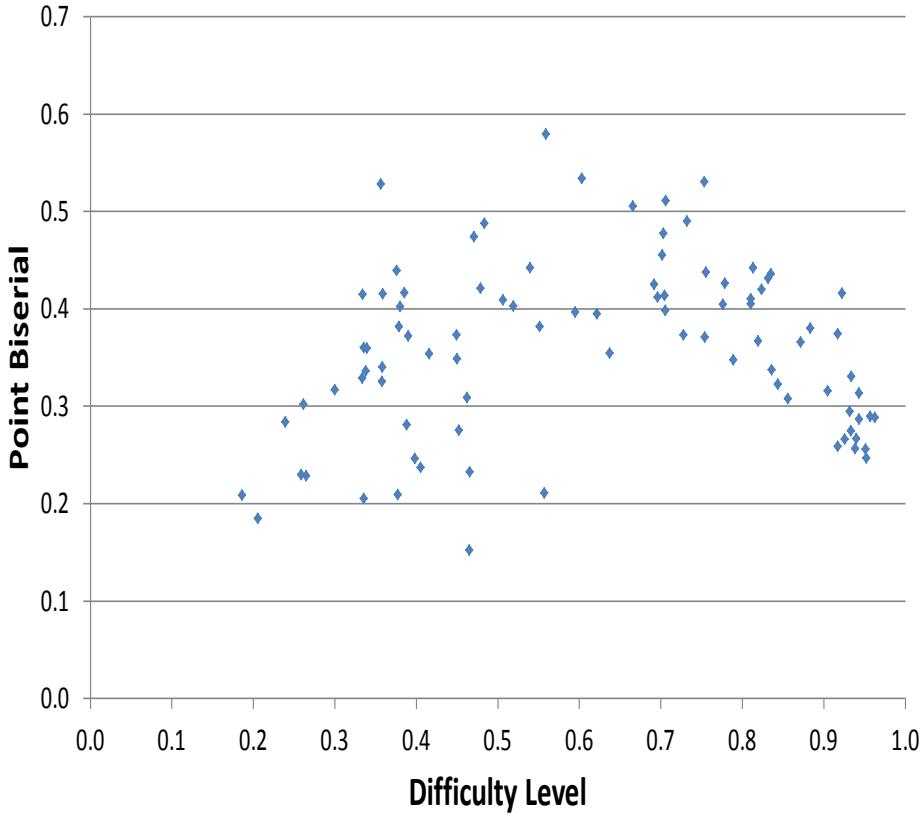






Slide 4

Classical Item Statistics (90 Items) (Assuming IRT Model Holds)



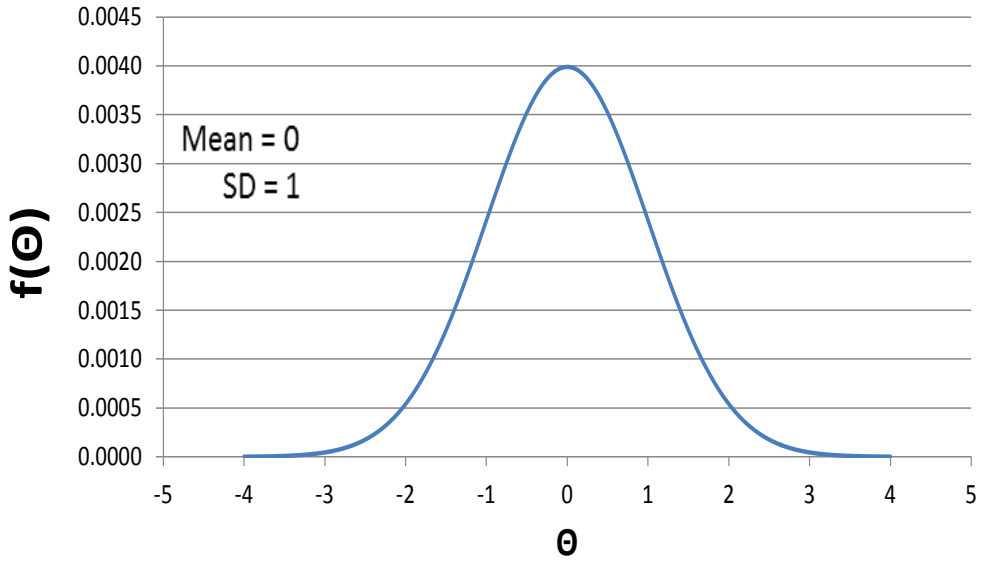
Slide 5

Parameter: Θ

Two Estimators:
Maximum Likelihood (MLE) and
Number Right (X)

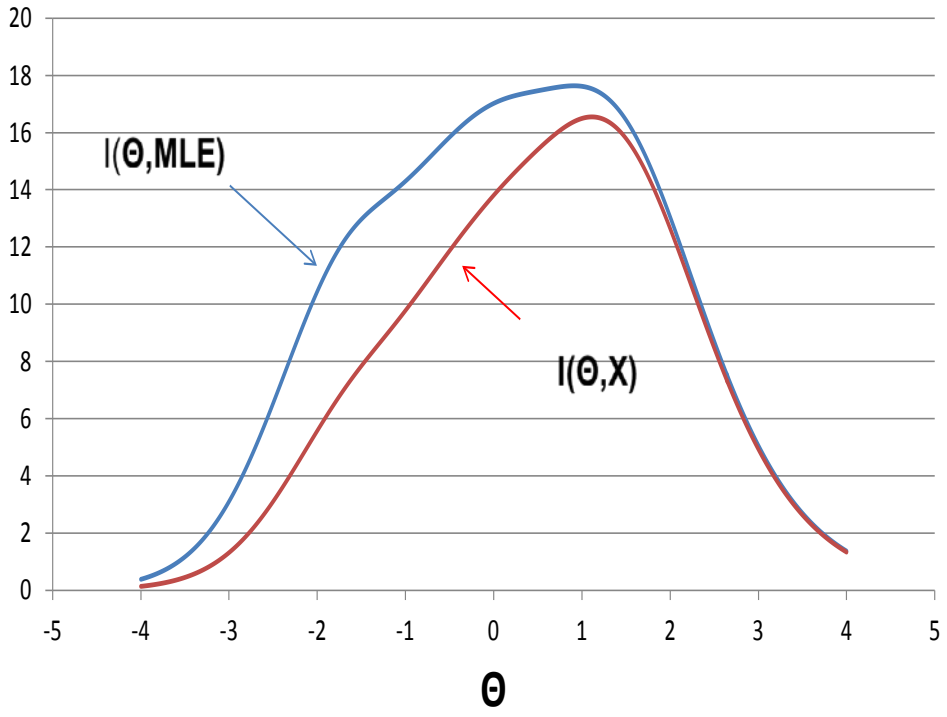
Slide 6

Distribution of Θ



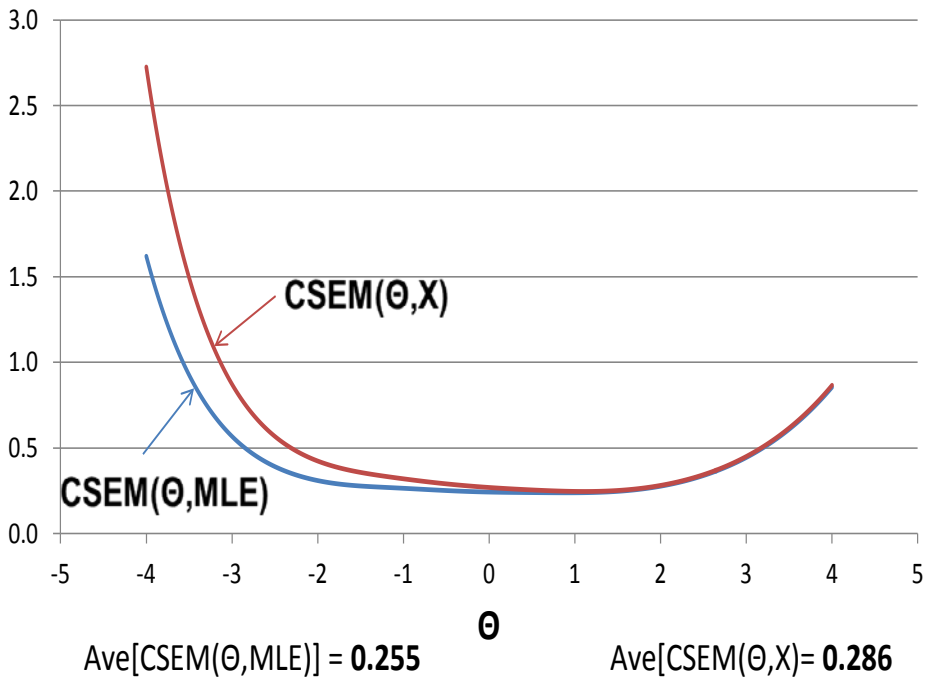
Slide 7

Information for Θ



Slide 8

Conditional Standard Errors of Measurement (CSEM) for Θ



Ave[CSEM(Θ,MLE)] = **0.255**

Ave[CSEM(Θ,X)= **0.286**

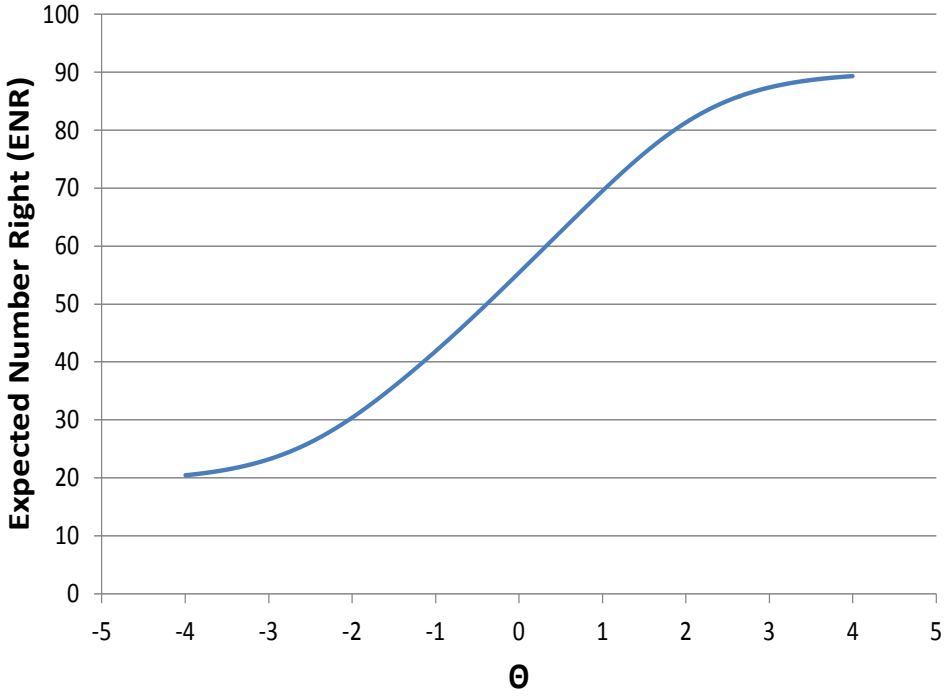
Slide 9

Parameter:
Transformation of Θ to
Expected Number Right (ENR)

Estimator:
Number Right (X)

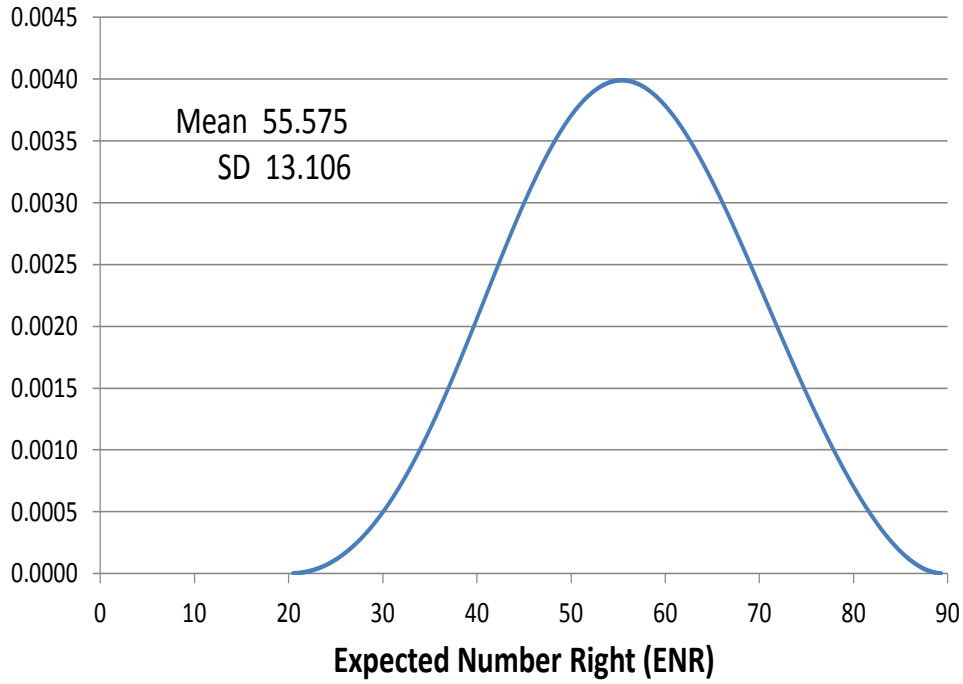
Slide 10

Test Characteristic Curve (TCC)



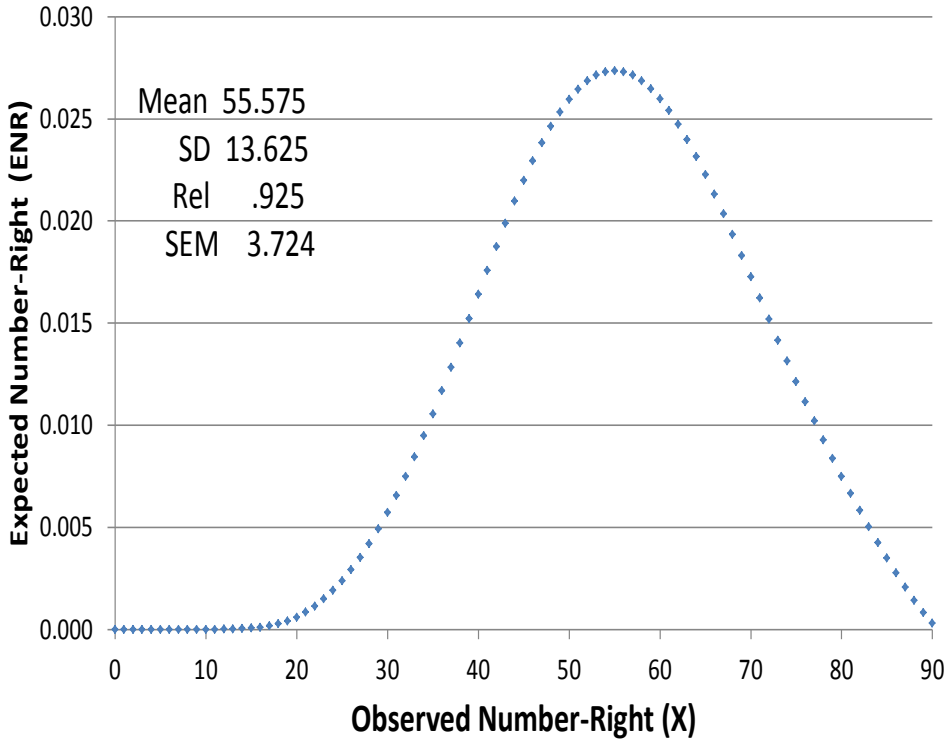
Slide 11

f(ENR) = True-score Distribution



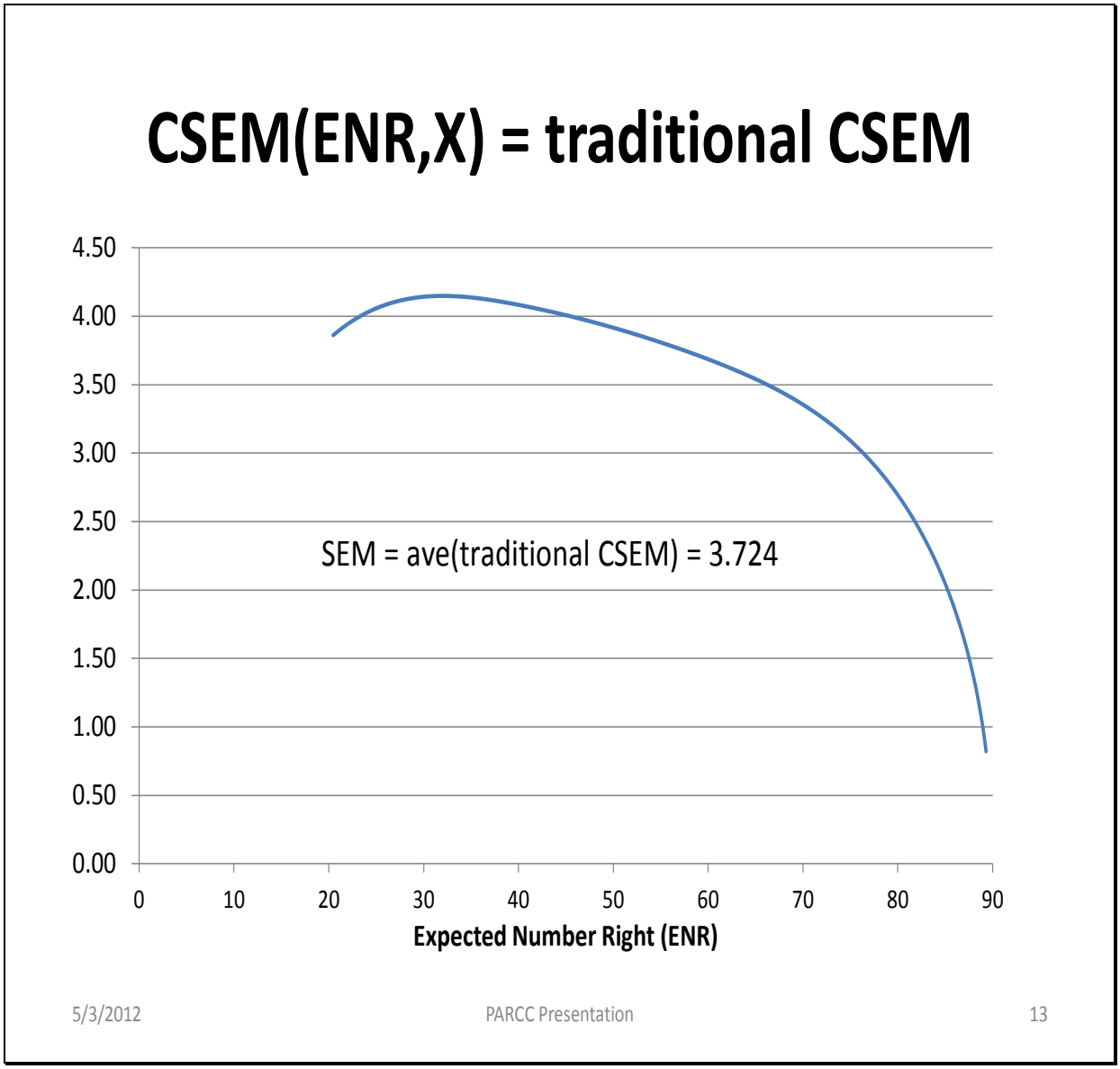
Slide 12

f(X) = Observed Number-Right Distribution



\

Slide 13


CSEM(ENR,X) = traditional CSEM
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
SEM = ave(traditional CSEM) = 3.724
0
10
20
30
40
50
60
70
80
90
Expected Number Right (ENR)
5/3/2012
PARCC Presentation
13

Slide 14

Parameter:

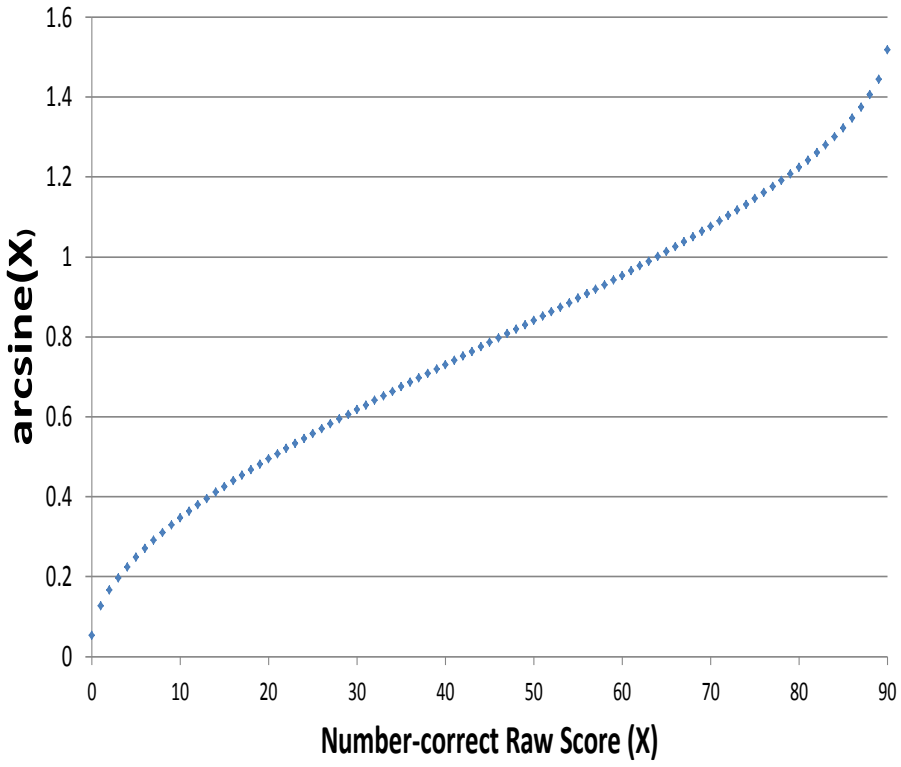
Arcsine Transformation of
Expected Number Right (ENR)
with an Additional Linear Transformation

Estimator:

Arcsine Transformation of X
with an Additional Linear Transformation

Slide 15

arcsine(x)
Purpose: Stabilize Error Variance



Slide 16

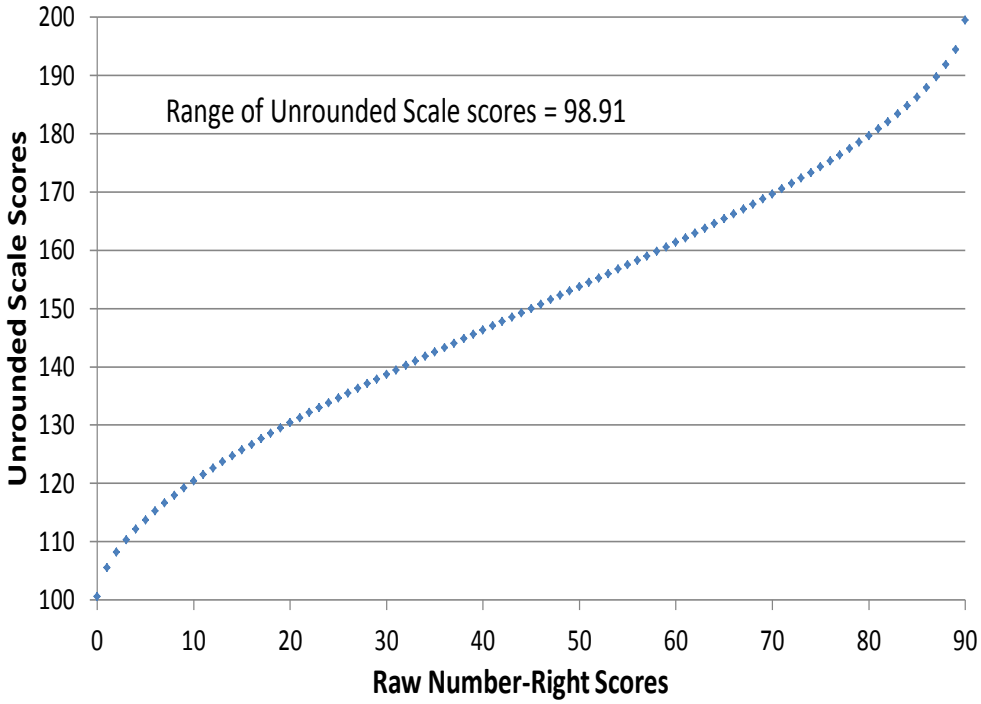
Let sc = scale score

GOAL: sc ± 3 covers true sc
68% of time for all sc (approximately)

sc = intercept + slope [arcsine(X)]
slope = 3/SD[arcsine(X)]
intercept = 150 – slope[arcsine(45)]

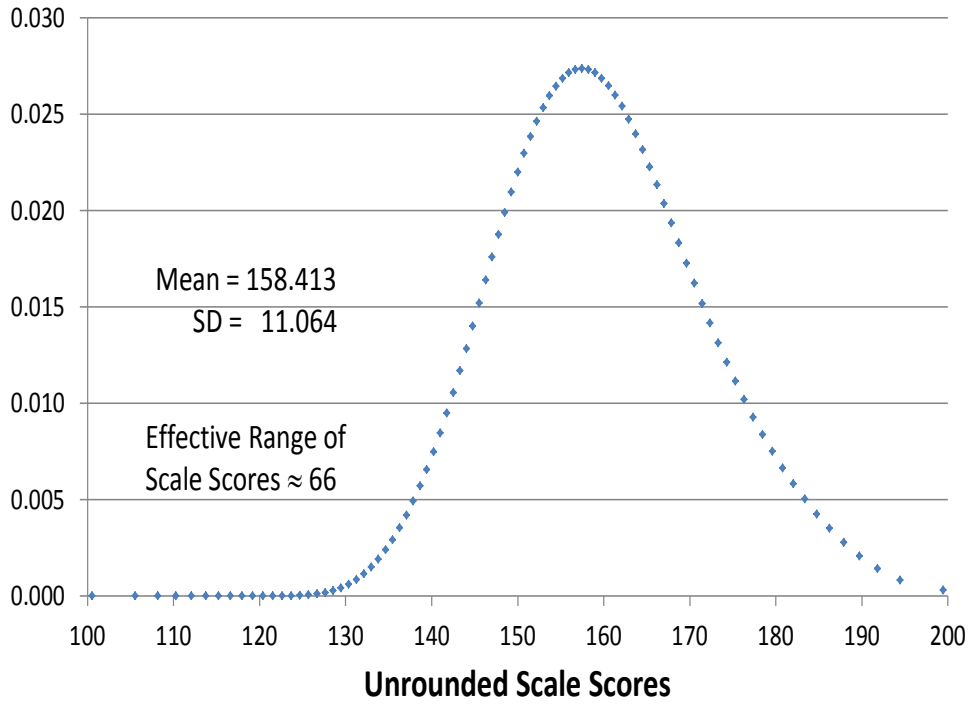
Slide 17

Conversion: Raw to Unrounded Scale Scores Using Arcsine with Ave. CSEM = 3 (arcsine(45) = 150)



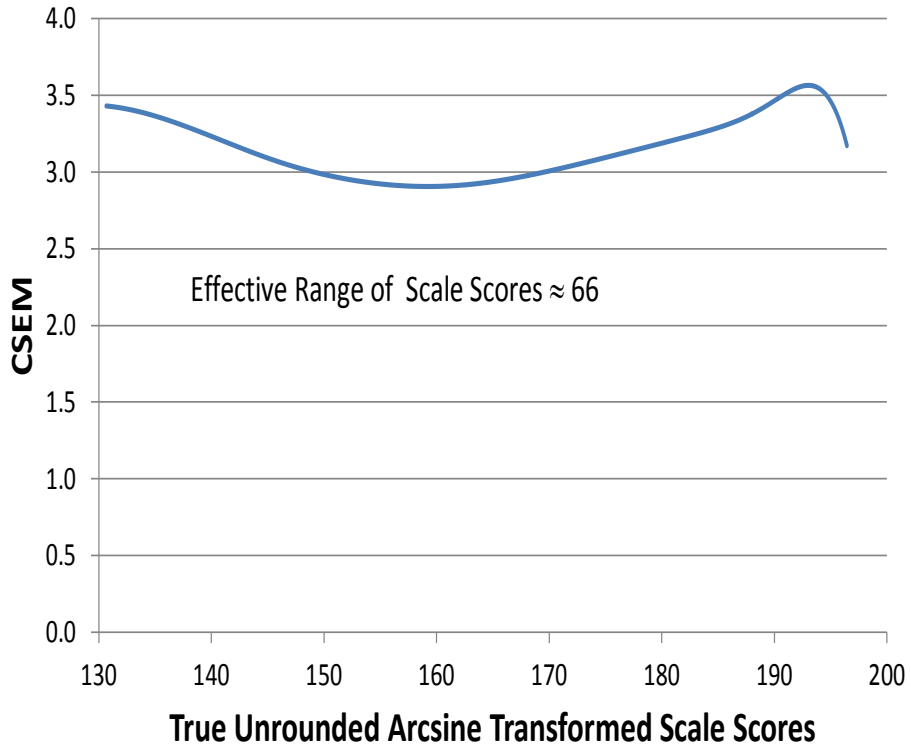
Slide 18

Relative Frequencies for Unrounded Scale Scores Using Arcsine with Ave. CSEM = 3 (arcsine(45) = 150)



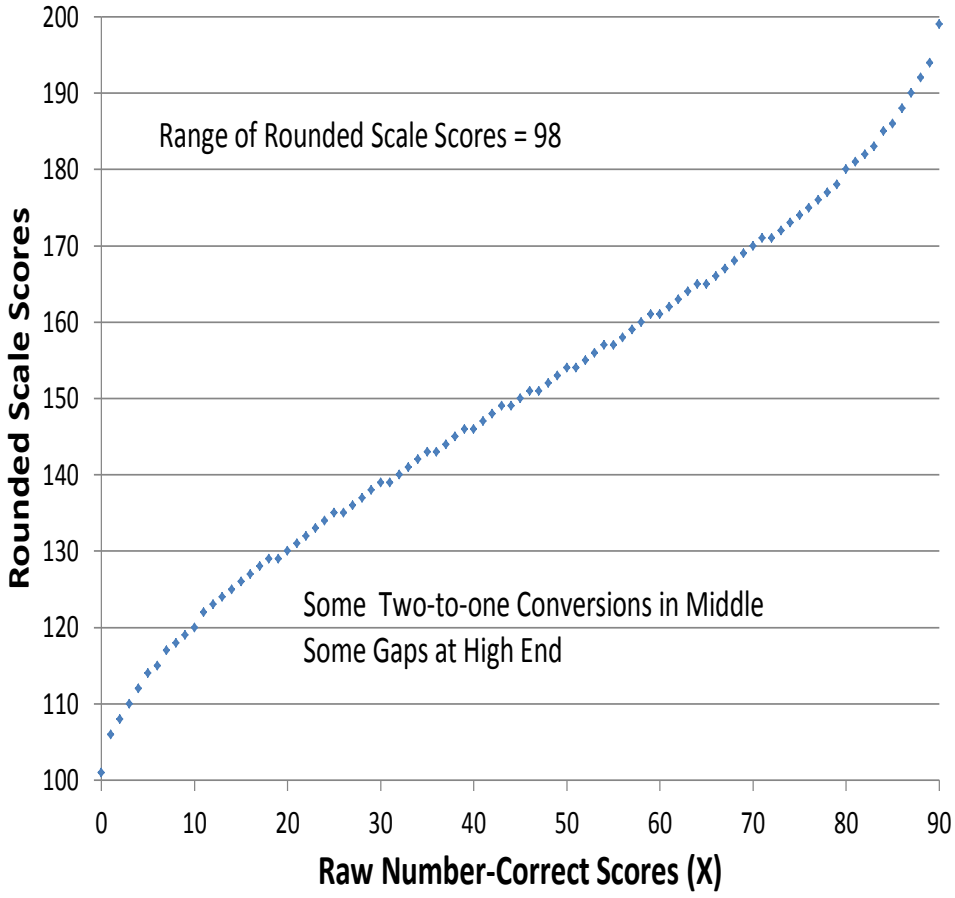
Slide 19

CSEM's for Unrounded Scale Scores Using Arcsine with Ave. CSEM = 3 (arcsine(45) = 150)

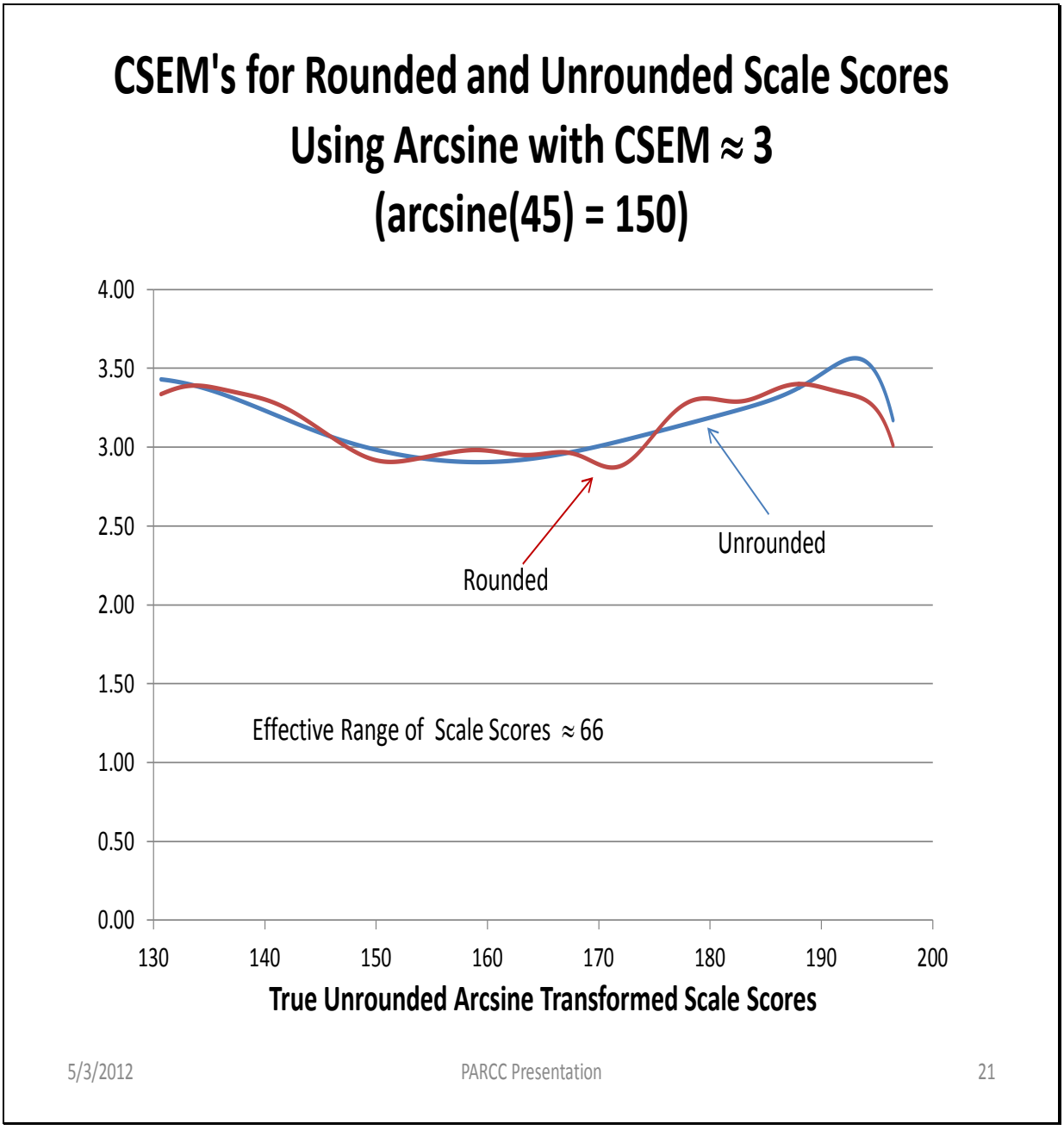


Slide 20

Conversion: Raw to Rounded Scale Scores
Using Arcsine with SEM = 3 and arcsine(45)=150



Slide 21


CSEM's for Rounded and Unrounded Scale Scores
Using Arcsine with CSEM ≈ 3
(arcsine(45) = 150)
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
130
140
150
160
170
180
190
200
Rounded
Unrounded
Effective Range of Scale Scores ≈ 66
True Unrounded Arcsine Transformed Scale Scores
5/3/2012
PARCC Presentation
21

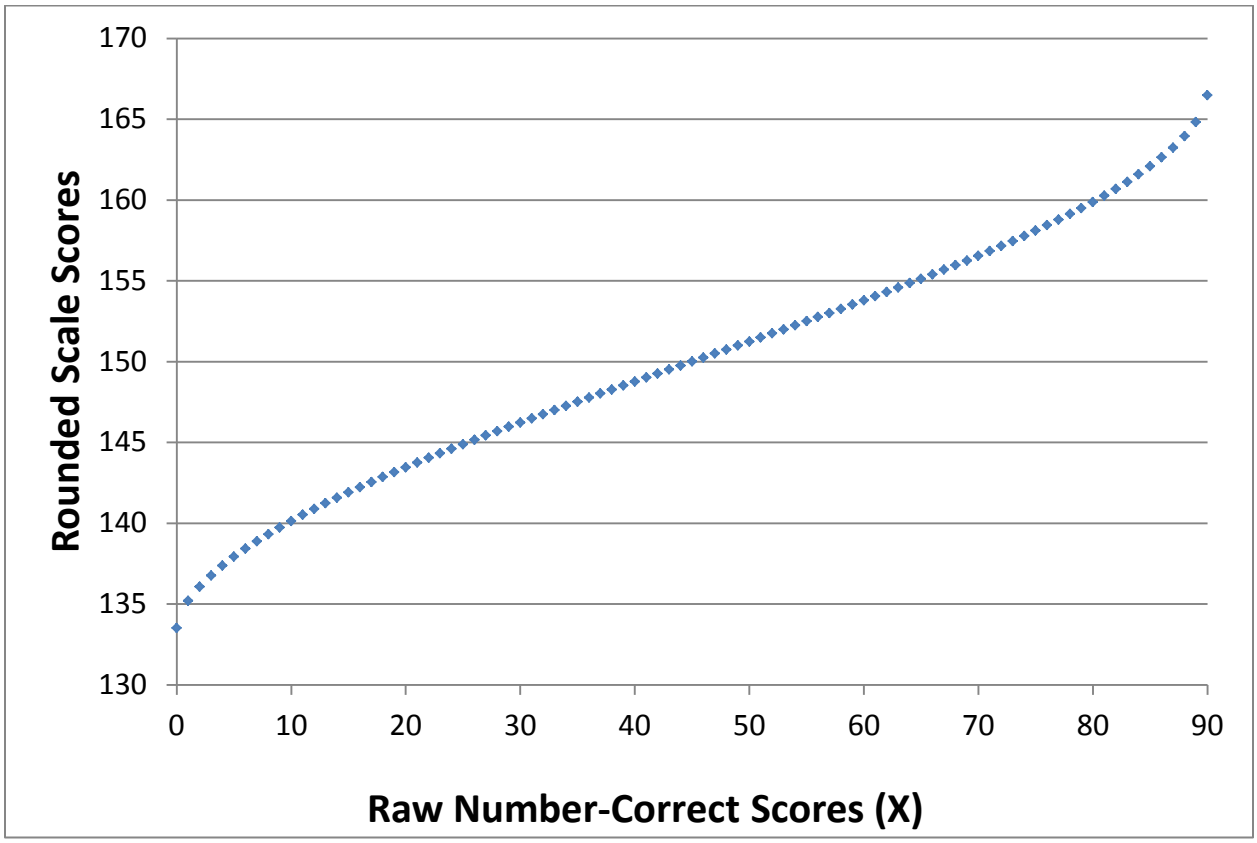
Slide 22

GOAL: sc ± 1 covers true sc

50% of time for all sc (approximately)

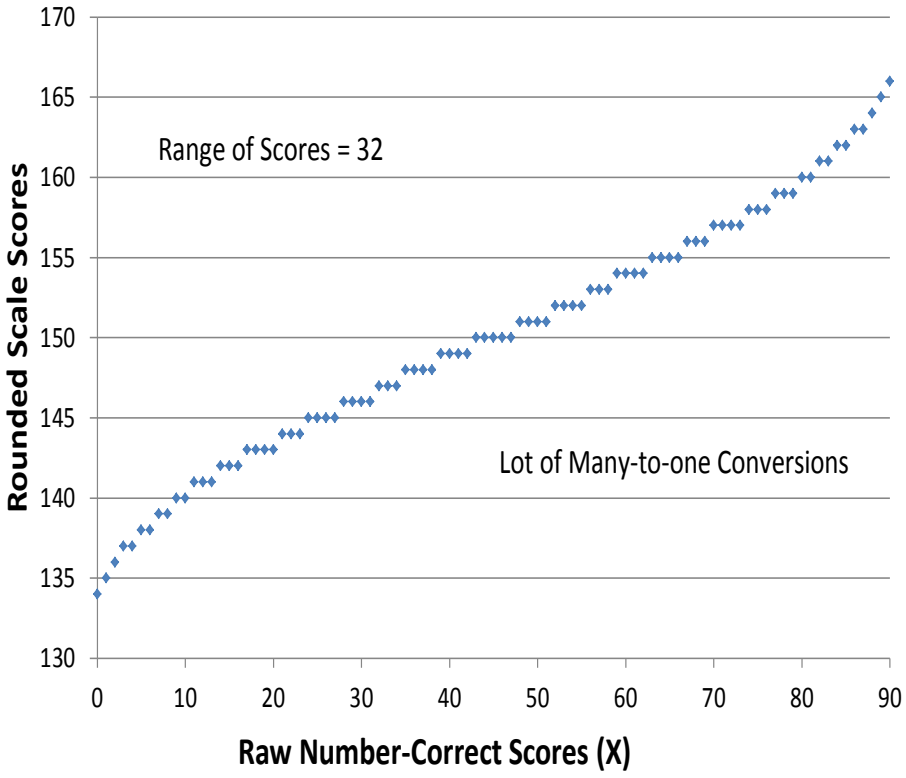
Slide 23

Conversion: Raw to Unrounded Scale Scores Using Arcsine with SEM = 1 and arcsine(45) = 150

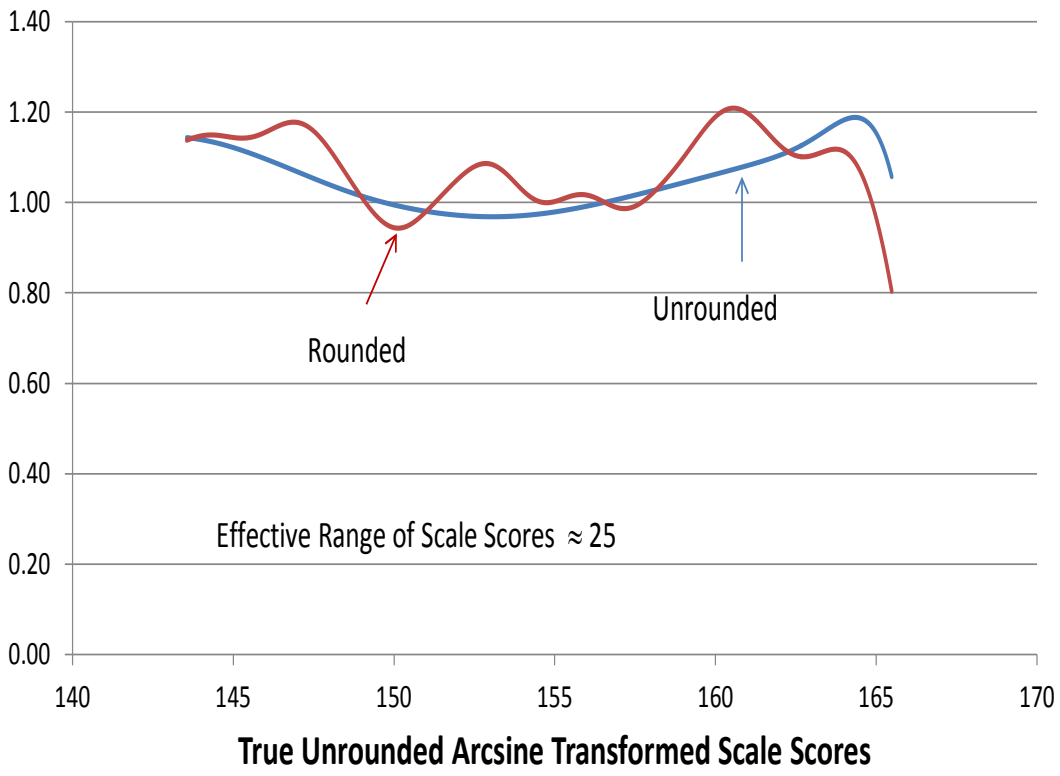


Slide 24

Conversion: Raw to Rounded Scale Scores
Using Arcsine with SEM = 1 and arcsine(45)=150



Slide 25

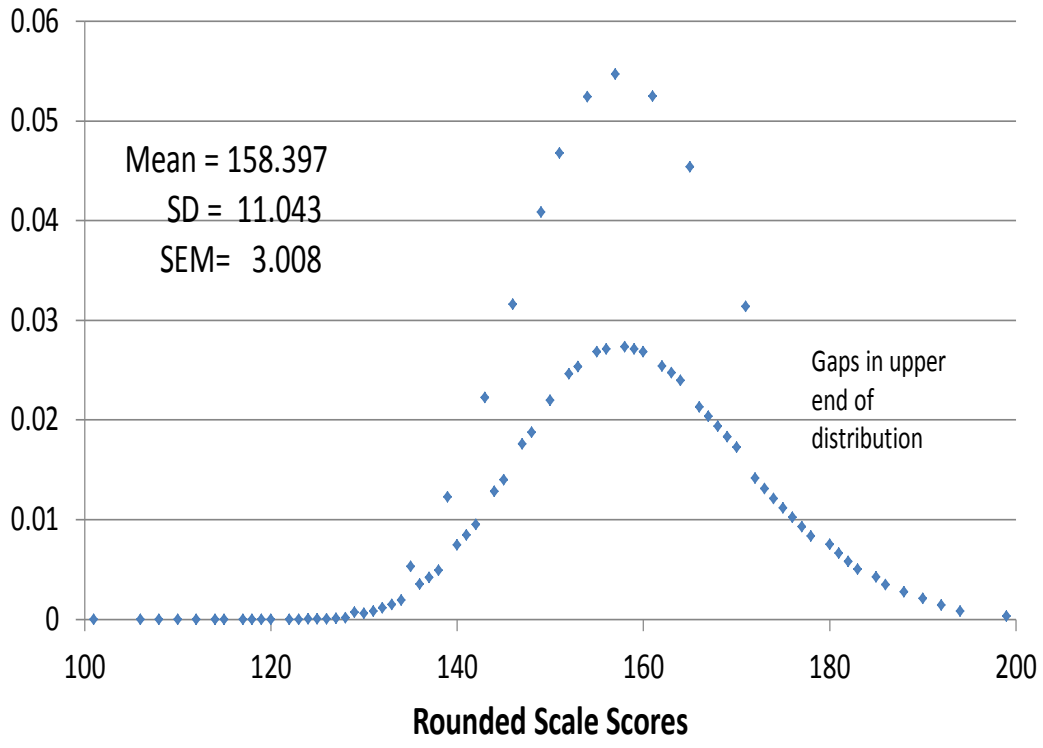

CSEM's for Rounded and Unrounded Scale Scores
Using Arcsine with CSEM ≈ 1 (arcsine(45) = 150)
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
Rounded
Unrounded
Effective Range of Scale Scores ≈ 25
140
145
150
155
160
165
170
True Unrounded Arcsine Transformed Scale Scores

Slide 26

Effects of Gaps and Many-to-one Conversions on Relative Frequencies of Rounded Scale Scores

Slide 27

Relative Frequencies for Rounded Scale Scores Using Arcsine with SEM=3 and arcsine(45) = 150



Slide 28

Relative Frequencies for Rounded Scale Scores Using Arcsine with SEM ≈ 1 and arcsine(45) = 150