Research Report

Aligning Scales of Certification Tests

Neil J. Dorans

Longjuan Liang

Gautam Puhan

February 2010

ETS RR-10-07

Aligning Scales of Certification Tests

Neil J. Dorans, Longjuan Liang, and Gautam Puhan
ETS, Princeton, New Jersey

February 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Dan Eignor

Technical Reviewers: Mary Grant and Tim Moses

Abstract

Scores are the most visible and widely used products of a testing program. The choice of score scale has implications for test specifications, equating, and test reliability and validity, as well as for test interpretation. At the same time, the score scale should be viewed as infrastructure likely to require repair at some point. In this report we examine the issue of scale fit—how well the scale fits the intended uses of its scores—for certification tests. Two examples of scale fit are considered: one in which the test has a single threshold that separates the candidate population into pass-fail groups, and one in which the test is required to support a restricted range of multiple thresholds.

Key words: scale alignment, score use, cut scores, certification tests

Table of Contents

List of Tables

Page

The choice of scales on which to report scores is one of a testing program's most fundamental and critical decisions. Scores are the most visible and widely used products of a testing program. They are what test takers get and what score users use. The score scale provides the framework for the interpretation of scores. The choice of score scale has implications for test specifications, equating, and test reliability and validity, as well as for test interpretation.

The utility of a score scale depends on how well it supports the inferences attached to its scores and how well it facilitates meaningful interpretations and minimizes misinterpretations (Petersen, Kolen, & Hoover, 1989). An important question, however, is how do we know that a score scale is serving its purpose properly? In this report we build upon score scale fit research conducted on the SAT® (Dorans, 2002) and the GRE® (Dorans, Yu, & Guo, 2006) and extend it to the domain of certification tests.

In section 1, we describe desirable scale properties that have been introduced and used elsewhere for broad-range tests such as admissions tests. In section 2, we introduce related scale properties for narrow-range tests such as certification exams, pointing out how they relate to the properties for scales of the broad-range tests. Section 3, which with section 2 is the core of the paper, contains measures of scale appropriateness that have been tailored to narrow-range tests. Section 4 presents an illustration involving a test with multiple cuts, while section 5 has an illustration of a single-cut exam. A discussion of scaling issues appears in section 6.

1. Desirable Scale Properties for Broad-Range Tests

What should a good scale look like? The scale should be well-aligned with the intended uses of the scores. For an assessment like the SAT, a broad-range test for which high, middle, and low scores may be pertinent for some admissions decisions, the degree to which the scale is well aligned depends on how the scale was originally defined and how well current score distributions fall on that scale. If scale alignment is desired for exams like the SAT, the well-aligned scale should possess several properties (Dorans, 2002).

The first desirable property of a broad-range test is that the scores of the reference group used to define the scale be *centered* near the midpoint of the scale. The average score (mean or median) in the reference group should be on or near the middle of the scale.

Second, the distribution of aligned scores for the scale-defining reference group should be *unimodal*, and that mode should be near the midpoint of the scale.

Third, the distribution should be nearly *symmetric* about the average score.

Fourth, the shape of the distribution should follow a commonly *recognized form*, such as the bell-shaped normal curve.

Fifth, the *working range* of scores should extend enough beyond the *reported range* of scores to permit shifts in population away from the scale midpoint without stressing the endpoints of the scale.

Sixth, the number of scale units should not exceed the number of raw score points, which is usually a simple function of the number of items. Otherwise, unjustified differentiation of examinees may occur.

Finally, a score scale should be viewed as infrastructure that is likely to require repair. Corrective action should be taken whenever average score distributions of current populations move sufficiently far away from the midpoint, or when distributions move far enough away from one of the endpoints to jeopardize the integrity of the scale at that endpoint.

The reasons for the first four properties are self-evident. If we want to maximize the longevity of the scale, we *center* the score distributions at the *center* of the score scale. Most human attributes have unimodal distributions. Given the symmetric or nearly symmetric nature of so many distributions of attributes, it seems logical to start with a symmetric distribution. The normal distribution is a unimodal symmetric distribution with a mathematically compact form that has known properties.

The fifth property allows the distribution of scores to shift over time until the situation is reached where the highest actual score is lower than the maximum reported score, or until the lowest actual score is higher than the minimum reported score. When the highest actual score falls short of the maximum reported score, then scores at the top end of the scale may be forced up to the maximum reported score via a scale-stretching process that may not produce exchangeable scores across editions of the test. As a result, scores may be misinterpreted. As in the case of the first property, having the working range subsume the score-reporting range allows a score scale to be useful for a longer time. The sixth property is the fundamental *one item, one scale point* property.[1] A gap[2] occurs when a one-point difference in raw scores translates to a multiple-point (2 or more) difference in scaled scores. A clump[3] occurs when two or more raw scores convert to the same scaled score. Gaps are worse than clumps.[4] Gaps exaggerate differences while clumps can hide them. To the extent that the score is unreliable, the exaggeration of differences is undesirable.

The placement of a unimodal, symmetric score distribution at the center of a reported score scale that is broad enough to accommodate shifts in the distribution should ensure that score interpretations are consistent and meaningful for an extended period of time. Provided the population of examinees is fairly stable, as is often the case with large populations, the score scale should be able to bear the subtle and slow-moving shifts in score distributions associated with that stable population.

2. Desirable Score Scale Properties for Narrow-Range Tests

For tests that separate the candidates into the proficient and non-proficient categories, desired score scale properties differ substantially from those cited above. The purpose of a certification test is to determine whether a candidate, such as a teacher, knows enough about his or her subject area to warrant certification. As far back as 1951, educational assessment professionals advocated constructing tests with difficulties aligned with the cut score or threshold used to distinguish between masters and non-masters. Davis (1951) suggested building a test in which "…every item should be as nearly as possible of 50 percent difficulty *for examinees at the level of ability represented by the 'passing mark.'*" (p. 318). Lord (1980), in his chapter on mastery testing, demonstrated how IRT can be used to achieve this goal by using more sophisticated indices such as test information curves and relative efficiencies. In this section we apply this threshold-focused thinking to the specification of desirable score scale properties for what we call narrow-range tests.

There are four desirable properties for scales associated with tests that focus on a narrow range of scores. The first desirable property of a narrow-range test is a scale centered on the cut score or threshold; if there are multiple thresholds for the test they should be equally distributed around the midpoint of the scale. This property explicitly recognizes the central importance of the cut score to the certification process. It is an adaptation of the first three principles cited above.

Second, the number of scale points should not exceed the number of raw score points. This is basically the same as the sixth desirable property of broad-range tests described earlier.

Third, scale construction needs to be sensitive to potential shifts in the difficulty of test forms and addition of new cut scores. The score scale range should be wide enough to accommodate shifts in test difficulty and the addition of new cut scores. This is an adaptation of the *working range* (fifth) principle cited earlier in the context of broad-range tests.

Finally, a score scale should be viewed as infrastructure likely to require repair. Corrective action should be taken whenever the cut scores are too close to one of the endpoints of the scale.

3. Measures of Scale Appropriateness

Several descriptive indices can be used in tandem to assess scale appropriateness for certification cases. The raw-to-scale conversion can be decomposed into several counts that describe its adequacy, namely the number of gaps, clumps, and one-to-one transformations. These indices also can be used for broad-range scales, not just narrow-range scales. In addition, another descriptive index, described below, can be used to quantify the portion of the score scale that is relevant to the intended use of the assessment.

For each raw score there is a corresponding scaled score that can be expressed in rounded or unrounded form. In this paper, we work with the unrounded conversions because they are not affected by the vagaries of rounding. In addition, we will work with the *essential score scale* (ESS). For example, in the case of a test like the SAT, the score reporting unit is 10. The essential score scale is 20 to 80 because the third digit on the 200-to-800 reported score scale is always zero; here the essential scale is obtained by dividing the reported score by 10. In general, a score scale can be converted to an essential score scale by dividing the upper and lower limits of the score scale by the score reporting unit, producing a scale with a reporting unit of 1.

Before defining the descriptive statistics that can be used to quantify the scale appropriateness, we will introduce the terms that will be used. Let ESS_k represent the unrounded essential scaled score associated with a raw score *k*. Let ESS_b represent the lowest unrounded scaled score that rounds to the bottom of the essential scaled score range, and ESS_t represent the highest unrounded score that rounds to the top of the essential scaled score range. For a 100-to-200 essential scaled score range, the lowest scaled score that rounded to 100 would be ESS_b, and the highest scaled score that rounded to 200 would be ESS_t. Let the corresponding raw scores be RS_b and RS_t, respectively. Let *m* equal the number of raw score points between RS_b and RS_t, inclusive.

If two consecutive raw scores yield essential scaled scores that are more that 0.5 units apart and less than 1.5 units apart, the essential transformation will be viewed as an *appropriate increment*. Otherwise it will be viewed as a *gap* or a *clump,* as defined here:

$$\begin{aligned} Gap_k &= 1 \leftrightarrow (ESS_k - ESS_{k-1}) \geq 1.5 \\ Clump_k &= 1 \leftrightarrow (ESS_k - ESS_{k-1}) \leq .5 \end{aligned} \tag{1}$$

The *GapCount* is defined as

$$GapCount = \sum_{b}^{t} Gap_k \tag{2}$$

while the *ClumpCount* is defined as

$$ClumpCount = \sum_{b}^{t} Clump_k . \tag{3}$$

There are $m - 1$ differences in essential scaled scores. Hence the number of appropriate increments is:

$$DesiredCount = m - 1 - ClumpCount - GapCount. \tag{4}$$

These three counts add up to $m - 1$; when divided by $m - 1$ they are expressed as percentages:

$$\begin{aligned} Desired\% &= DesiredCount/(m-1) \\ Clump\% &= ClumpCount/(m-1) \\ Gap\% &= GapCount/(m-1) \end{aligned} \quad . \tag{5}$$

The final index to be introduced, RPR (relevant proportion ratio), is the proportion of the score scale that is relevant to the certification purpose of the assessment. This index is determined in three steps. The hardest step is the first. Using the standard error of judgment (SEJ)[5] from the standard setting that produced the cut score(s) on the test, we can place a confidence band around the raw cut score(s). For the test with a single cut, we simply use plus or minus three SEJs around the raw cut score. For the test with multiple cut scores, the band would be (lowest raw cut score – 3 SEJ, highest raw cut score + 3 SEJ). Next these boundaries are converted to the essential score scale, and these unrounded scores are rounded to the nearest integer on the essential score scale to produce ESS_l for the lower boundary and ESS_u for the upper boundary. Finally, the ratio

$$RPR = (ESS_u - ESS_l)/(ScaleRange) \quad (6)$$

defines the portion of the scaled score range that is relevant to the certification purposes of the assessment. *Scale Range* would be 200 – 100 = 100 for a 100-to-200 essential scaled score range. The larger the RPR, the better the scale is for its intended uses.

In the next two sections we apply these indices to score scales of two testing programs, one which employs multiple cut scores (testing program A) and the other which employs a single cut score (testing program B). For both testing programs we identified two exams each, one where there is approximately a one-to-one correspondence between raw and scaled score and the other where there is approximately a one-to-two correspondence between raw and scaled scores. Then we suggest an alternative scaling that increases the RPR and minimizes the *ClumpCount* and the *GapCount* in the relevant region of the score scale.

4. Illustration with Test Data: Multiple Cut

Table 1 is a summary of the gapping and clumping associated with four forms of each of two different certification exams from testing program A. One exam typically contains 50 items; the other exam typically contains 120 items. This table shows

- The actual number of items on each form.
- The number of possible points on the reported score scale.
- One less than the minimum number of raw score points needed to span the reported score range ($m - 1$).
- Three count measures that sum to $m - 1$: *GapCount* (Equation 2), *ClumpCount* (Equation 3), and *DesiredCount* (Equation 4).
- The percent versions of these counts (divide by $m - 1$) — $Gap\%$, $Clump\%$, $Desired\%$ (all in Equation 5).
- The RPR (Equation 6). (There are no RPRs for the four forms of Exam 2 are not included in the table because the original SEJs were not available for these forms.)

Table 1

Summary of Scale Appropriateness Indices for Two Exam Titles From Testing Program A

Exam	Form	No. of items	No. of scale points[a]	$m-1$	Counts			Percentage			RPR
					Gaps	Clumps	Desired	Gaps	Clumps	Desired	
1	1	49	101	34	34	0	0	100.00	0	0	0.57
	2	50	101	35	35	0	0	100.00	0	0	0.54
	3	49	101	35	35	0	0	100.00	0	0	0.55
	4	50	100	35	35	0	0	100.00	0	0	0.54
2	1	117	101	79	7	0	72	8.86	0	91.14	-
	2	118	101	75	16	0	59	21.33	0	78.67	-
	3	119	101	85	0	1	84	0	1.18	98.82	-
	4	119	101	85	0	0	85	0	0	100.00	-

Note. This statistic is not available for Exam 2 because the original standard error of judgment (SEJ) is not available. m = the minimum number of raw score points needed to span the entire reported score range. RPR = relevant proportion ratio.

[a] The conversions for Forms 1, 2, and 3 of Exam 1 all range from 100 to 200 (which makes 101 scale points), while the conversion for form 4 ranges from 100 to 199 (which makes 100 scale points).

As an example consider the measures of form appropriateness calculated for Exam 1, Form 1. As seen in Table 2, the number of points between rounded scaled scores 200 and 100 is 35 (i.e., $m = 35$). The difference between two successive unrounded scaled scores is considered a gap if greater than 1.5. As seen in Table 2, the differences between two successive unrounded scores for the essential score region are all greater than 1.5, leading to 34 gaps (i.e., 34 gap counts in Table 1). Since the difference between two successive unrounded scores for the entire essential score region is greater than 1.5, the clump count automatically becomes zero for the essential score region (i.e., clump counts in Table 1). The percentages of *GapCount*, *ClumpCount*, and *DesiredCount* in Equation 4 are calculated by dividing gap count and clump counts, from Equation 4, by $m - 1$. Finally, the RPR of 0.57 is obtained by taking the difference between the highest and lowest essential rounded scaled score in Table 2 (169 - 112 = 57) and dividing it by 100 (which is the scale score range).

Table 2

Existing Raw-to-Scale Conversion for Form 1 of Exam 1 From Testing Program A

Raw	Unrounded	Rounded	$ESS_k - ESS_{k-1}$	Gap	Clump
0	65.666	100			
1	68.536	100			
2	71.405	100			
3	74.275	100			
4	77.144	100			
5	80.014	100			
6	82.883	100			
7	85.753	100			
8	88.622	100			
9	91.492	100			
10	94.361	100			
11	97.231	100			
12	100.101	100			
13	102.970	103	2.870	1	0
14	105.840	106	2.870	1	0
15	108.709	109	2.870	1	0
16	111.579	112	2.869	1	0
17	114.448	114	2.870	1	0
18	117.318	117	2.870	1	0
19	120.187	120	2.870	1	0
20	123.057	123	2.870	1	0
21	125.926	126	2.870	1	0
22	128.866	129	2.940	1	0
23	131.804	132	2.938	1	0
24	134.739	135	2.935	1	0
25	137.651	138	2.912	1	0
26	140.548	141	2.897	1	0
27	143.425	143	2.877	1	0
28	146.284	146	2.859	1	0
29	149.128	149	2.845	1	0
30	151.956	152	2.828	1	0
31	154.779	155	2.822	1	0
32	157.590	158	2.811	1	0
33	160.404	160	2.815	1	0
34	163.230	163	2.826	1	0
35	166.100	166	2.870	1	0
36	168.969	169	2.870	1	0
37	171.839	172	2.869	1	0
38	174.708	175	2.870	1	0
39	177.578	178	2.870	1	0
40	180.447	180	2.869	1	0

Raw	Unrounded	Rounded	$ESS_k - ESS_{k-1}$	Gap	Clump
41	183.317	183	2.870	1	0
42	186.187	186	2.870	1	0
43	189.056	189	2.870	1	0
44	191.926	192	2.869	1	0
45	194.795	195	2.869	1	0
46	197.665	198	2.870	1	0
47	200.534	200			
48	203.404	200			
49	206.273	200			

Note. The lightly shaded rows highlight the raw score range that matters. The darkest row marks the center of the existing scale. ESS_k-ESS_{k-1} = the essential scale score range the difference between the unrounded scale score in row k and the unrounded scale score in row k-1.

Note that the gapping problem is more problematic for the exam with fewer test items, given the same length of essential score range. In contrast the percentage of adequate raw-to-scale mappings is generally higher than 90% for the longer test. Form 2 of Exam 2 is a noticeable exception, but its 21% $Gap\%$ is far superior to the 100% noted for all four forms of Exam 1.

4.1 Form 1 of Exam 1 in More Detail

Form 1 of Exam 1 is used to illustrate the setup of the new scale. As noted, forms in this exam usually contain 50 multiple choice questions. One item in Form 1 was identified as a problem item and did not count towards the total score. Thus the maximum possible raw score point is 49, as seen in the top row of Table 1, which also contains a summary of the indices that measure the scale appropriateness for this particular form.

Different user states adopted this test for certification purposes, and each state has its own cut score on a reference test form. The lowest cut score among all the states that adopted this test is 123, and the highest is 156 on the 100-to-200 scale. These cut scores were obtained by transforming the raw cut scores recommended using a modified Angoff standard setting method to scale score units using the raw-to-scale conversion for that reference test form. The Angoff method is the most widely used and thoroughly researched standard-setting method applied to certification testing (see Brandon, 2004; Hurtz & Hertz, 1999). The SEJ for this test is 1.3 in raw score points and will be used to define the new score scale.[6]

The existing raw-to-scale conversion is represented in Table 2, which contains unrounded and rounded scaled scores for each raw score between 0 and 49. Column $ESS_k - ESS_{k-1}$ calculates the difference between two adjacent unrounded essential scaled scores. If this difference is greater than or equal to 1.5, the column *Gap* will have a value of 1, indicating a gap; otherwise it has a value of 0. If this difference is smaller than or equal to 0.5, column *Clump* will have a value of 1, indicating a clump; otherwise it will have a value of 0. Note that scores above 200 and below 100 are ignored.

This scale violates two principles of desirable score scale properties for a narrow-range test described in section 2. First of all, the number of scaled score points (101) exceeds the number of raw score points (49). In Table 2 we see that a 1-point difference in raw score results in 2 or 3 points in the rounded scaled score. This creates a lot of gaps. In Table 2, we also see that the differences between two adjacent scaled scores are all greater than 1.5. The $Gap\%$ is 100%. There are no clumps for this scale. The $Desired\%$ of raw to-scale conversions is 0.

Note that this scale also violates the first principle "the scale should be centered on the cut score or threshold." The lightly shaded region in Table 2 highlights the raw score range that matters. The darkly shaded row is the center of the existing scale. Note that although a scaled score of 150 should be the center of a 100-to-200 scale, the current raw-to-scale conversion table did not have a scaled score of 150; therefore a scaled score of 152, the next higher score above 150, was considered as the center of the scale. Clearly the center of the scale is not the center of the raw score range that matters in practice. The number of scaled score points that may be relevant to certification purposes is 57; over 40% of the scale is clearly irrelevant.

4.2 Improving the Relevance of the Exam 1 Scale

In this section we describe the steps used to build a new scale possessing the desirable properties for a narrow-range test that we have described in section 2. We use Form 1 of Exam 1 to set up this scale.

1. *Identify the raw score range on Form 1 that matters*. The raw score minimum (RS_l) for Form 1 is defined as three SEJs below the raw cut score for the state with the lowest cut score. The raw score maximum (RS_u) for Form 1 is defined as three SEJs above the raw cut score for the state with the highest cut score. For Form 1 the lowest raw cut score is 20 (which corresponds to the smallest scaled cut score of 123), and the highest raw cut

score is 32 (which corresponds to the highest scaled cut score of 156). The SEJ is 1.3. Thus $RS_u = 36$ and $RS_l = 16$.

2. *Map the midpoint of the scale (ESSmid) to the midpoint of this raw score range that matters*. For this example the midpoint of a 100–200 scale is 150. The midpoint of the raw score range that matters is 26, halfway between 16 and 36. Thus for the new scale a raw score of 26 will have a scaled score of 150.
3. *Establish a relationship such that a 1-point increase in raw scores results in a 1-point increase in scaled scores for the raw score range that matters*. In other words, the slope of the raw-to-scale conversion should be 1, and the intercept would depend on where we wanted to place this one-to-one relationship on the scale. As indicated in step 2 above that would be at the midpoint. However, we propose a raw-to-scale relationship whereby a 1-point increase in raw scores results in a slightly less than a 1-point increase in scaled scores. The slope and intercept of such a relationship is defined as

$$A = \frac{RS_u - RS_l}{(RS_u + 0.499) - (RS_l - 0.5)}$$
$$B = ESS_{mid} + (RS_u - RS_l)/2 - A \times (RS_u + 0.499). \quad (7)$$

This particular relationship is used, instead of the simpler version that has A = 1 and does not have -0.5 or 0.449, to delay the eventual development of gaps in the score scale that can result from chains of subsequent equating. The slope and intercept are obtained by treating the integers RS_l and RS_u as if they are continuous. So the lower end of R_l, ($R_l - 0.5$), and the upper end of RS_u, $(RS_u + 0.499)$, are used. Use of 0.499 and -0.5 should increase the shelf life of the new scale.

4. *Truncate raw scores outside the range that is relevant to the range of cut scores.* Any raw score lower than RS_l has a scaled score which is the same as the rounded scaled score corresponding to RS_l. Any raw score higher than RS_u has a scaled score which is the same as the rounded scaled score corresponding to RS_u.

Applying these steps to Form 1, which had the scaling properties depicted in the top row of Table 1 and in greater detail in Table 2, yields a revised scale that is shown in Table 3.

Table 3

Revised Scale for Form 1 of Exam 1 From Testing Program A

Raw score	Unrounded scaled score	Rounded scaled score	$ESS_k - ESS_{k-1}$	Gap	Clump
0	140.000	140			
1	140.000	140			
2	140.000	140			
3	140.000	140			
4	140.000	140			
5	140.000	140			
6	140.000	140			
7	140.000	140			
8	140.000	140			
9	140.000	140			
10	140.000	140			
11	140.000	140			
12	140.000	140			
13	140.000	140			
14	140.000	140			
15	140.000	140			
16	140.476	140			
17	141.429	141	0.952	0	0
18	142.381	142	0.952	0	0
19	143.333	143	0.952	0	0
20	144.286	144	0.952	0	0
21	145.238	145	0.952	0	0
22	146.191	146	0.952	0	0
23	147.143	147	0.952	0	0
24	148.096	148	0.952	0	0
25	149.048	149	0.952	0	0
26	150.000	150	0.952	0	0
27	150.953	151	0.952	0	0
28	151.905	152	0.952	0	0
29	152.858	153	0.952	0	0
30	153.810	154	0.952	0	0
31	154.763	155	0.952	0	0
32	155.715	156	0.952	0	0
33	156.667	157	0.952	0	0
34	157.620	158	0.952	0	0
35	158.572	159	0.952	0	0
36	159.525	160	0.952	0	0

Raw score	Unrounded scaled score	Rounded scaled score	$ESS_k - ESS_{k-1}$	Gap	Clump
37	160.000	160			
38	160.000	160			
39	160.000	160			
40	160.000	160			
41	160.000	160			
42	160.000	160			
43	160.000	160			
44	160.000	160			
45	160.000	160			
46	160.000	160			
47	160.000	160			
48	160.000	160			
49	160.000	160			

Note. The lightly shaded rows highlight the raw score range that matters. The darkest row marks the center of the existing scale. ESS_k-ESS_{k-1} = the essential scale score range the difference between the unrounded scale score in row k and the unrounded scale score in row k-1.

Table 3 shows that the revised scale meets all desirable properties proposed for certification tests. The center of the scale now is also the center of the score range that matters. The number of scale points is the same as the number of raw score points that may be relevant to certification. No gaps or clumps are present.

Our third desirable property for narrow-range tests is that the scale construction be sensitive to potential shifts in the difficulty of test forms and the addition of new cut scores. Table 4 summarizes how the conversion line changes after a few equatings. In Table 4, Form 1 is the revised scale described in Table 3. Operationally, Form 2 was equated to Form 1, Form 3 to Form 2, and Form 4 to Form 3. It is clear from Table 4, that after a few equatings there are no gaps present in any of the conversions. After a few equatings, however, a few clumps occur at both tails of the score scale. For example, on Form 4 two clumps occur at raw scores of 16 and 17, and two other clumps occur at raw scores of 42 and 43. These clumps were caused by the truncation of the scales at both ends, and these raw scores points are not critical in making pass/fail decisions.

Table 4

Scale Appropriateness Indices Across Several Forms of Exam 1 From Testing Program A

Exam	Form	No. of items	No. of scale points	$m-1$	Counts			Percentage			RPR
					Gaps	Clumps	Desired	Gaps	Clumps	Desired	
1	1	49	21	20	0	0	20	0	0	100.00	100
	2	50	21	23	0	1	22	0	4.35	95.65	100
	3	49	21	24	0	2	22	0	8.33	91.67	95
	4	50	21	27	0	4	23	0	14.81	85.19	95

Note. RPR = relevant proportion ratio. m = the minimum number of raw score points needed to span the entire reported score range.

This section illustrates how well aligned the existing 100-200 score scale from a certification testing program was to a short test and to a long test. Table 1 demonstrated that the short tests containing approximately 50 items are not served well by a score scale that contains 101 points. In addition, the nature of the existing scaling results in large regions of the score scale that are used but that are not relevant to the goal of certification. In sum, certification purposes for the short test are not served well by a score scale that exaggerates small differences between raw scores. The longer exam, on the other hand, had adequate scale properties with respect to gaps and clumps. Without an SEJ, however, it was not possible to calculate the RPR. Hence it was hard to estimate what proportion of the reported score range was actually relevant to certification decisions.

5. Illustration With Test Data: Single Cut

The exams used in this section have only a single cut score and scores that range from 100 to 300. Hence its score scale can be more focused. Table 5 summarizes how well aligned the score scales of two exams are with their intended uses. It contains a summary of the gapping and clumping associated with three forms each of two different exams from testing program B. One exam contains 80 items; the other exam contains 180 items. This table also shows:

- The actual number of items on each form.
- The number of possible points on the reported score scale.
- The minimum number of raw score points needed to span the reported score range ($m-1$).
- Three count measures that sum to $m-1$—*GapCount* (Equation 2), *ClumpCount* (Equation 3), and *DesiredCount* (Equation 4).

- The percent versions of these counts (divide by $m - 1$) — *Gap*%, *Clump*%, *Desired*% (all in Equation 5), and the RPR (Equation 6). (There are no RPRs for the three forms of Exam 2 included in this table because the original SEJs were not available for these forms.)

We see again that the shorter test exhibited 100% gaps and lacked any desired raw-to-scale mappings. In contrast, the longer tests exhibited no gapping or clumping. However, two forms did not scale out to the full range of the score scale, as evidenced by number of scale points less than 201 for Forms 2 and 3.

The RPRs for the first exam range from 14 to 17. These small numbers mean that between 80% and 85% of the score scale range, 100 to 300, has no relevance to certification. The RPRs for the longer test could not be calculated.

5.1 Form 1 of Exam 1 in More Detail

Form 1 of Exam 1 is used to illustrate the new way to set up the scale. As noted, forms in this test usually contain 80 multiple-choice questions. The first row of Table 5 contains a summary of the indices that measure the scale appropriateness for this particular form.

The existing raw-to-scale conversion is represented in Table 6, which contains unrounded and rounded scaled scores for each raw score between 0 and 80. Form 1 is the very first form of this exam title.

Table 5

Summary of Scale Appropriateness Indices for Two Exams From Testing Program B

Exam	Form	No. of items	No. of scale points	$m - 1$	Counts			Percentage			RPR
					Gaps	Clumps	Desired	Gaps	Clumps	Desired	
1	1	80	201	80	80	0	0	100	0	0	17
	2	80	179	80	80	0	0	100	0	0	15
	3	80	169	80	80	0	0	100	0	0	14
2	1	180	201	175	0	0	175	0	0	100	-
	2	180	182	180	0	0	180	0	0	100	-
	3	180	189	173	0	0	173	0	0	100	-

Note. m = the minimum number of raw score points needed to span the entire reported score range. RPR = relevant proportion ratio. This statistic is not available for Exam 2 because the original standard error of judgment (SEJ) is not available.

Table 6

Existing Raw-to-Scale Conversion for Form 1 of Exam 1 From Testing Program B

Raw score	Unrounded scaled score	Rounded scaled score	$ESS_k - ESS_{k-1}$	Gap	Clump	Raw score	Unrounded scaled score	Rounded scaled score	$ESS_k - ESS_{k-1}$	Gap	Clump
0	100.000	100				41	217.143	217	2.8572	1	0
1	102.857	103	2.8571	1	0	42	220.000	220	2.8571	1	0
2	105.714	106	2.8572	1	0	43	222.857	223	2.8571	1	0
3	108.571	109	2.8571	1	0	44	225.714	226	2.8572	1	0
4	111.429	111	2.8572	1	0	45	228.571	229	2.8571	1	0
5	114.286	114	2.8571	1	0	46	231.429	231	2.8572	1	0
6	117.143	117	2.8572	1	0	47	234.286	234	2.8571	1	0
7	120.000	120	2.8571	1	0	48	237.143	237	2.8572	1	0
8	122.857	123	2.8571	1	0	49	240.000	240	2.8571	1	0
9	125.714	126	2.8572	1	0	50	241.936	242	1.9355	1	0
10	128.571	129	2.8571	1	0	51	243.871	244	1.9355	1	0
11	131.429	131	2.8572	1	0	52	245.807	246	1.9355	1	0
12	134.286	134	2.8571	1	0	53	247.742	248	1.9354	1	0
13	137.143	137	2.8572	1	0	54	249.677	250	1.9355	1	0
14	140.000	140	2.8571	1	0	55	251.613	252	1.9355	1	0
15	142.857	143	2.8571	1	0	56	253.548	254	1.9355	1	0
16	145.714	146	2.8572	1	0	57	255.484	255	1.9355	1	0
17	148.571	149	2.8571	1	0	58	257.419	257	1.9355	1	0
18	151.429	151	2.8572	1	0	59	259.355	259	1.9354	1	0
19	154.286	154	2.8571	1	0	60	261.290	261	1.9355	1	0
20	157.143	157	2.8572	1	0	61	263.226	263	1.9355	1	0
21	160.000	160	2.8571	1	0	62	265.161	265	1.9355	1	0
22	162.857	163	2.8571	1	0	63	267.097	267	1.9355	1	0
23	165.714	166	2.8572	1	0	64	269.032	269	1.9355	1	0
24	168.571	169	2.8571	1	0	65	270.968	271	1.9354	1	0
25	171.429	171	2.8572	1	0	66	272.903	273	1.9355	1	0

Raw score	Unrounded scaled score	Rounded scaled score	$ESS_k - ESS_{k-1}$	Gap	Clump	Raw score	Unrounded scaled score	Rounded scaled score	$ESS_k - ESS_{k-1}$	Gap	Clump
26	174.286	174	2.8571	1	0	67	274.839	275	1.9355	1	0
27	177.143	177	2.8572	1	0	68	276.774	277	1.9355	1	0
28	180.000	180	2.8571	1	0	69	278.710	279	1.9355	1	0
29	182.857	183	2.8571	1	0	70	280.645	281	1.9355	1	0
30	185.714	186	2.8572	1	0	71	282.581	283	1.9354	1	0
31	188.571	189	2.8571	1	0	72	284.516	285	1.9355	1	0
32	191.429	191	2.8572	1	0	73	286.452	286	1.9355	1	0
33	194.286	194	2.8571	1	0	74	288.387	288	1.9355	1	0
34	197.143	197	2.8572	1	0	75	290.323	290	1.9355	1	0
35	200.000	200	2.8571	1	0	76	292.258	292	1.9355	1	0
36	202.857	203	2.8571	1	0	77	294.194	294	1.9354	1	0
37	205.714	206	2.8572	1	0	78	296.129	296	1.9355	1	0
38	208.571	209	2.8571	1	0	79	298.065	298	1.9355	1	0
39	211.429	211	2.8572	1	0	80	300.000	300	1.9355	1	0
40	214.286	214	2.8571	1	0						

Note. The lightly shaded rows highlight the raw score range that matters. The darkest row marks the center of the existing scale. ESS_k-ESS_{k-1} = the essential scale score range the difference between the unrounded scale score in row k and the unrounded scale score in row k-1.

The original scale was set by mapping the raw cut score to a scaled cut score of 240. The line for scores above the raw cut score was obtained by mapping the maximum raw score point with the maximum scaled score point, which is 300. The line for scores below the raw cut score point was obtained by mapping the lowest raw score point (which will be 0) with the minimum scaled score point, which was 100. Column $ESS_k - ESS_{k-1}$ in Table 6 calculates the difference between two adjacent unrounded essential scaled scores. If this difference is greater than or equal to 1.5, the column *Gap* will have a value of 1, indicating this is a gap; otherwise it has a value of 0. If this difference is smaller than or equal to 0.5, column *Clump* will have a value of 1, indicating this is a clump; otherwise it will have a value of 0.

This scale violates two of the principles of a desirable score scale for a narrow-range test that we have recommended. First, the number of scaled score points (201) exceeds the number of raw score points (80). In Table 6, we see that a 1-point difference in raw score results in 2 or 3 points of difference in the rounded scaled score. This creates a number of large gaps. In Table 6, we also see that the differences between two adjacent scaled scores are all greater than 1.5. They are 2.85 for raw scores of 0 to 49 and 1.94 for raw score of 50 to 80. The *Gap*% is 100%. There are no clumps for this scale. The *Desired*% of raw-to-scale conversions is 0.

Note also that this scale also violates the first principle that the scale should be centered on the cut score or threshold. The cut score is 240, which corresponds to a raw score of 49. The SEJ for this test was 2.2 in raw score points. Hence the region that matters for certification is 42 to 56, 49 plus or minus (3 × 2.2) in rounded score units. The lightly shaded region in Table 6 highlights this raw score range that matters. The darkly shaded row marks the center of the scale. Clearly the center of the scale is not the center of the raw score range that matters in practice. In fact the center of the scale, 200, is below the minimum scaled score that matters of 220. The number of scaled score points that may be relevant to certification purposes is 17; over 80% of the scale is clearly irrelevant to the intended purpose of the test.

5.2 Improving the Relevance of the Exam 1 Scale

In this section we apply the steps used to build a new scale described in section 4.2. We use Form 1 of Exam 1 to set up this scale and follow the steps previously listed:

1. Identify the raw score range on Form 1 that matters.

2. Map the midpoint of the scale (ESS_{mid}) to the midpoint of this raw score range that matters.
3. Establish a relationship such that a 1-point increase in raw scores results in a 1- point increase in scaled scores for the raw score range that matters.
4. Truncate raw scores outside the range that is relevant to the range of cut scores.

Applying these steps to Form 1, which had the scaling properties depicted in the top row of Table 5 and in greater detail in Table 6, yields the revised scale given in Table 7.

Table 7 shows that the revised scale meets all desirable properties proposed for certification tests. The center of the scale now is also the center of the score range that matters. The number of scale points is the same as the number of raw score points that may be relevant to certification. No gaps and no clumps are present.

The third desirable property for narrow-range test scales in section 2 states that the scale construction needs to be sensitive to potential shifts in the difficulty of test forms and addition of new cut scores. Table 8 summarizes how the conversion line changes after a few equatings. In Table 8, Form 1 is the revised scale described in Table 7. Operationally, Form 2 was equated to Form 1, and Form 3 was equated to Form 2. It is clear from Table 8 that after a few equatings there are no gaps present in any of the conversions. After a few equatings, however, a few clumps occur at both tails of the score scale, where raw scores points are not critical in making pass/fail decisions.

This section illustrates how well aligned the existing 100-to-300 score scale for testing program B was to a short test and to a long test. Table 5 demonstrated that the short tests that contain approximately 80 items are not served well by a score scale that contains 201 points. In addition, the nature of the existing scaling results in very large regions of the score scale that are used but which are not relevant to the goal of certification. In sum, certification purposes of the short test are not served well by a score scale that exaggerates small differences between raw scores.

The longer exam, on the other hand, had adequate scale properties with respect to gaps and clumps. Without a SEJ, it was not possible, however, to calculate the RPR. Hence it was hard to estimate what proportion of the reported score range was actually relevant to certification decisions.

Table 7

Revised Scale for Form 1 Exam 1 From Testing Program B

Raw score	Unrounded scaled score	Rounded scaled score	$ESS_k - ESS_{k-1}$	Gap	Clump	Raw score	Unrounded scaled score	Rounded scaled score	$ESS_k - ESS_{k-1}$	Gap	Clump
0	193.000	193				41	193.000	193			
1	193.000	193				42	193.467	193			
2	193.000	193				43	194.400	194	0.933	0	0
3	193.000	193				44	195.333	195	0.933	0	0
4	193.000	193				45	196.267	196	0.933	0	0
5	193.000	193				46	197.200	197	0.933	0	0
6	193.000	193				47	198.134	198	0.933	0	0
7	193.000	193				48	199.067	199	0.933	0	0
8	193.000	193				49	200.000	200	0.933	0	0
9	193.000	193				50	200.934	201	0.933	0	0
10	193.000	193				51	201.867	202	0.933	0	0
11	193.000	193				52	202.801	203	0.933	0	0
12	193.000	193				53	203.734	204	0.933	0	0
13	193.000	193				54	204.667	205	0.933	0	0
14	193.000	193				55	205.601	206	0.933	0	0
15	193.000	193				56	206.534	207	0.933	0	0
16	193.000	193				57	207.000	207			
17	193.000	193				58	207.000	207			
18	193.000	193				59	207.000	207			
19	193.000	193				60	207.000	207			
20	193.000	193				61	207.000	207			
21	193.000	193				62	207.000	207			
22	193.000	193				63	207.000	207			
23	193.000	193				64	207.000	207			
24	193.000	193				65	207.000	207			
25	193.000	193				66	207.000	207			
26	193.000	193				67	207.000	207			

Raw score	Unrounded scaled score	Rounded scaled score	$ESS_k - ESS_{k-1}$	Gap	Clump	Raw score	Unrounded scaled score	Rounded scaled score	$ESS_k - ESS_{k-1}$	Gap	Clump
27	193.000	193				68	207.000	207			
28	193.000	193				69	207.000	207			
29	193.000	193				70	207.000	207			
30	193.000	193				71	207.000	207			
31	193.000	193				72	207.000	207			
32	193.000	193				73	207.000	207			
33	193.000	193				74	207.000	207			
34	193.000	193				75	207.000	207			
35	193.000	193				76	207.000	207			
36	193.000	193				77	207.000	207			
37	193.000	193				78	207.000	207			
38	193.000	193				79	207.000	207			
39	193.000	193				80	207.000	207			
40	193.000	193									

Note. The lightly shaded rows highlight the raw score range that matters. The darkest row marks the center of the existing scale. ESS_k-ESS_{k-1} = the essential scale score range the difference between the unrounded scale score in row k and the unrounded scale score in row k-1.

Table 8

Scale Appropriateness Indices Across Several Forms of Exam 1 From Testing Program B

Exam	Form	No. of items	No. of scale points	$m - 1$	Counts			Percentage			RPR
					Gaps	Clumps	Desired	Gaps	Clumps	Desired	
1	1	80	15	14	0	0	14	0	0	100.00	100.00
	2	80	15	17	0	0	17	0	0	100.00	85.71
	3	80	15	19	0	2	17	0	10.53	89.47	85.71

Note. RPR = relevant proportion ratio. m = the minimum number of raw score points needed to span the entire reported score range.

6. Discussion

6.1 The Existing Scale Situation

All scales are arbitrary, but not all scales are equally useful. We have listed four principles that describe properties that a good score scale should have if it is to be used for certification purposes. None of the scales examined in this study are appropriate for the intended uses of the score, namely certification. One obvious problem is that for some tests, the number of items is smaller than the number of scale score points. The existence of exaggerated raw-to-scale differences for these short tests probably results in unfairness for examinees having proficiencies near the cut score region who happened to take a form that was slightly easier in the cut score region than another version of the test. If they had been given the harder form they may have had a better chance of demonstrating their minimal competence. For example, assume the cut score is 50, and that one edition of a test has reported scores of 49 and 52, while another edition of the test has reported scores of 48 and 51. Here the 3-point gap matters. Examinees right at the cut are more likely to pass if they take the form with the reported score sequence of 48 and 51 (most likely score) than if they take the form with the reported score sequence of 49 (most likely score) and 52. This is not an equating issue—an equating based on 1 million examinees cannot overcome the unfairness associated with gaps in score conversion tables. The shorter tests from these two testing programs are prone to this kind of scaling unfairness, but the longer tests are not.

All tests we have considered have existing scales with large portions that are irrelevant to the intended purpose of the test. Form 1 of Exam 1 from testing program B illustrates this point

clearly—the only region of interest was between 220 and 254, which means all scores below 220 (121 score points) and above 254 (46 score points) are outside the region of interest. This waste of space has consequences for test assembly and efficiency of measurement, as discussed in the next section.

There are concerns with truncating the wasted space, however. Candidates who exceed their state's passing score know what they need to know – they passed. While some of these successful candidates may have interest in supplemental information, the majority will be satisfied to know they passed. Those who did not pass probably want to know by how much they missed passing. This fact may argue against truncating the lower scores.

A long scale with much wasted space is not the only means of meeting the needs of those who failed to pass the test. The number of items by which they missed the cut score could be reported to them. This number would be purely descriptive. In addition, they could be told the probability that they would pass if they took a test very similar to it soon and without any additional study or preparation. Livingston and Lewis (1995) described one approach to determining this probability. Both the number of questions short of passing and the probability statement about how likely they would be to pass upon immediate retesting could be used to supplement the primary score scale that is centered around the range of cut scores and which does not contain scores that fall far outside this restricted region of interest.

6.2 Implications for Test Assembly

The present scales for the two testing programs are longer than they need to be, given their intended use. As a consequence, measurement power is spread out over a wide range of scores, some of which are of little practical relevance to certification. A test that is shorter but comprised of items that have maximum measurement power near the cut scores might serve certification purposes better. The scales that we proposed in the illustrations above were centered around the cut score(s) of interest. The test should be comprised of items that are expected to measure well in these regions. As noted earlier, educational assessment professionals for a long time have advocated constructing tests with difficulties that are aligned with the cut score or threshold used to distinguish between masters and non-masters. Very easy items serve little purpose, other than allowing examinees to warm up to the tasks ahead. Very difficult questions are even less useful; when they are extremely difficult a large proportion of the few who answer them correctly may have done so more because of luck than skill.

Lord (1980), in his chapter on mastery testing, demonstrated how item response theory (IRT) can be used to achieve this goal of targeting tests towards selected regions of the score scale. He used sophisticated indices such as test information curves and relative efficiencies. An early example of this can be found in Dorans and Livingston (1983). They used IRT tools to develop test specifications for the PSAT/NMSQT®, a broad-range test. This use of IRT to assemble the tests is a sensible practice, especially if equating is conducted as a check on the assembly process.

IRT tools can be used to guide the test assembly process for a certification test. Focusing a scale around the cut points of interest, as described in this paper, produces a target scale for IRT or other test assembly tools. If the easy and hard items were to be trimmed away and replaced by a smaller number of more appropriately difficult items, the length of the operational test would be shortened without any loss of efficacy. More testing time would become available for pretesting items. Over time, the item writers may become better at writing items that measure best in the scaled score region of interest.

In sum, by aligning the score scale with the intended uses of the scores, fewer items will be needed for assessment, the items that are needed will be more targeted to a specific portion of the score scale than items that comprise the current tests, and testing time can be used to collect pretest information for assembling more focused tests of higher quality.

7. Summary

Several simple ways of summarizing score fit have been introduced, using simple counting rules. One set of counts provides an indication of how far the raw-to-scale conversion used to set the scale deviates from the ideal of a one-to-one relationship between raw scores and reported scores. The second indicates what portion of the score scale is germane to the intended use of the test to certify candidates. These counting rules can be improved upon to distinguish small gaps from larger gaps. Illustrations were drawn from the two certification programs, and new types of scales are proposed. Implementation of these scales should lead to improved certification decisions, and more focused and less expensive assessment.

References

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*, 59-88.

Davis, F. B. (1951). Item selection techniques. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 266-328). Washington, DC: American Council on Education.

Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: How and why. *Journal of Educational Measurement, 39*(1), 59-84.

Dorans, N. J., & Livingston, S. A. (1983). *Statistical specifications for the PSAT/NMSQT: Recommendations for the present and alternatives for the future* (ETS Statistical Rep. No. SR-83-66). Princeton, NJ: ETS.

Dorans, N. J., Yu, L., & Guo, F. (2006). *Evaluating scale fit for broad-ranged admissions tests* (ETS Research Memorandum No. RM-06-04). Princeton, NJ: ETS.

Hurtz, G. M. & Hertz, N. R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. *Educational and Psychological Measurement, 59*, 885-897.

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

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179-197.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York, NY: Macmillan.

Notes

[1] Alternative approaches, as well as alternative prescriptions, to determining the number of scaled score points on a test are discussed in Chapter 9 of Kolen and Brennan (2004).

[2] As an example of gapping, consider the SAT scale prior to around 1970. At that time, any whole number between 200 and 800 was a possible score, and one additional correct answer could raise the scaled score by 8 or more points. One version of a test may have reported scores of 631, 639, 648, while another version may have reported scores of 633, 641, 650. These differences, for example of 639 vs. 641, could have led a user to think that the scores were more precise than they actually were. If 640 were a critical score for decision-making purposes, the students who took the second test would have been advantaged. On a scale with only two moving digits, both tests would have reported scores of 630, 640, and 650.

[3] As an example of clumping, consider the AP® (Advanced Placement®) exam. Here raw scores can take on many possible values, but scores are reported on a 5-point scale—1, 2, 3, 4, and 5. Consequently, many test takers with very different raw scores can receive the same scaled score. In addition, a few test takers with very similar raw scores can receive different scaled scores due to rounding.

[4] At first glance, clumps appear to be worse than gaps because potentially useful information is lost when two raw scores convert to the same scaled score. On any single test form this loss of information is undesirable and so gapping, which does not discard information, may be viewed as preferable to clumping. This picture changes, however, when scores are compared across different forms of the same test, as shown in the example in Note 2.

[5] The SEJ is a measure of the extent to which the recommended cut scores would vary if the standard setting were replicated with many different samples of judges.

[6] Note that this SEJ is from the standard setting done for the first form. Form 1 in Table 1 is not the first form of this exam title. However, we will still use this SEJ for illustration purpose.