Accuracy of Year-1, Year-2 Comparisons Using Individual Percentile Rank Scores: Classical Test Theory Calculations

David Rogosa
Stanford University
August 1999

National Center for Research on
Evaluation, Standards, and Student Testing

Deliverable - August 1999

Project 3.3 Validity of Interpretations and
Reporting of Results-- Measurement of Progress

Project Director: David Rogosa

U.S. Department of Education
Office of Educational Research and Improvement
Award #R305B60002

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles

Accuracy of Year-1, Year-2 Comparisons Using Individual Percentile Rank Scores: Classical Test Theory Calculations

David Rogosa
Stanford University
July 1999

ABSTRACT

In the reporting of individual student results from standardized tests in educational assessments, the percentile rank of the individual student is a major, if not the most prominent, numerical indicator. For example, in the 1998 and 1999 California Standardized Testing and Reporting (STAR) program using the Stanford Achievement Test Series, Ninth Edition, Form T (Stanford 9), the 1998 *Home Report* and 1999 *Parent Report* feature solely the National Grade Percentile Ranks. (These percentile rank scores also featured in the more extensive *Student Report*). This paper develops a formulation and presents calculations to examine the properties of year-1, year-2 comparisons using these individual percentile rank scores. The approach and formulation follows the previous investigations of the accuracy of the individual percentile rank score in Rogosa (1999a). A typical question that this paper addresses is: What are the chances that a student who really improved 10 percentile points from year-1 (1998) to year-2 (1999) obtains a lower percentile rank in year-2 than year-1? Such questions are addressed using the test reliability coefficient in classical test theory to represent quality of measurement. Thus we can investigate the question, What level of test reliability is needed to obtain good accuracy in year-1, year-2 comparisons?

1. Technical Formulation

The technical formulation is a basic errors-in-variables model with all components having Gaussian distributions. As in Rogosa (1999a), all that is meant by the phrase “classical test theory calculation” is to identify the calculations as pertaining to the simplest case of constant error variance across the score distribution with continuous, Normally distributed scores. The components of what is referred to as the classical test theory calculation are listed below; use the subscript i = 1,2 to indicate properties of year-1 or year-2 measurements.

- The cumulative distribution function of the observed scores Y in the national norming sample is denoted by $G_i(Y)$ for year-1 or year-2. The classical test theory formulation defines this norming distribution, with density function $g_i(Y)$, to be a Normal Distribution; denote the corresponding population mean and standard deviation for Y by (μ_{Ni} , σ_{Ni}) for year-1 or year-2.

- The observed measure Y contains error of measurement ε . The *classical test theory* assumptions dictate that the error of measurement, denoted by ε, has a Normal Distribution with mean 0 and constant variance $\sigma^2_{\varepsilon i}$ across the score distribution: i.e., $\varepsilon \sim N(0, \sqrt{\sigma^2_{\varepsilon i}}\ )$. (More general formulations, such as σ^2_{ε} depending on the level of the test score, can be incorporated into many of these results, with the overhead of added complexity.)

- The test reliability coefficient is often used as an index of the quality of measurement. The test reliability is defined for the full (norms) population; from the *classical test theory* formulation, the reliability is $rel_i = (\sigma^2_{Ni} - \sigma^2_{\varepsilon i}\ )/\sigma^2_{Ni}$. For a rough, but useful, illustration set the reliability of a 60-item test to be .90 (in line with standardized achievement tests). Then use Spearman-Brown to obtain the rough test length equivalents for various reliability values:

reliability	.60	.65	.70	.75	.80	.85	.90	.95
number items	10	12	16	20	27	38	60	127

- The norming distributions, $G_i(Y)$, are based on fallible Y-scores. An alternative is to consider what the norming distribution would be if measurement had been perfect (i.e., not distorted by error of measurement in Y). At the risk of over-complicating the notation, denote by $G^*_i(Y)$ the cumulative distribution function with corresponding mean and standard deviation (μ_{Ni} , $[(\sigma^2_{Ni} - \sigma^2_{\varepsilon i}\ )]^{½}$); $G^*_i(Y)$ represents a (hypothetical) norming distribution not distorted by measurement error (i.e., constructed from scores with reliability 1).

- The score for an individual student examinee is denoted by S, The percentile rank (PR) for the score S is 100 $G_i(S)$; thus $G_i(S)$ can be thought of as a nondecreasing transformation of the score S to the percentile rank metric. The score S has underlying true score τ ; the measurement model is $S = \tau + \varepsilon$. An individual under perfect measurement has percentile rank in observed norming distribution 100 $G_i(\tau)$ or, in a norming distribution not distorted by measurement error, the percentile rank is 100 $G^*_i(\tau)$. Often in the calculations, an individual (or an individual's achievement level) is characterized a value of $G^*_i(\tau)$.

2. Accuracy of Year-to-Year Improvement in Percentile Rank Scores

The observed improvement is $G_2(S_2) - G_1(S_1)$, the signed difference between the percentile rank scores for year-2 and year-1; improvement may be positive or negative. The main accuracy calculation is for the quantity:

$$y1y2 = \Pr\{\, G_2(S_2) - G_1(S_1) \ \le\ \text{bound} \mid G_1^*(\tau_1),\ G_2^*(\tau_2)\} \quad , \qquad (1)$$

the probability that the improvement in the percentile rank scores is less than or equal to the quantity "bound" (bound may be negative or positive) for a student with stated year-1 and year-2 values $G_1^*(\tau_1)$ and $G_2^*(\tau_2)$.

2.1 Computation of y1y2 Probability: Technical Details

Preliminaries. Let $\Phi[x]$ indicate the distribution function (cdf) for N(0,1) and $\phi[x]$ indicate the density (pdf) for N(0,1). Then $G_i(x) = \Phi[(x - \mu_{Ni})/\sigma_{Ni}]$, and $S_i \mid \tau_i \sim N[\,\tau_i\,, \sigma_{Ni}(1 - \text{rel}_i)^{1/2}\,]$ so that $\Pr\{\, S_i \le\ x\} = \Phi[(x - \tau_i)/\sigma_{Ni}(1 - \text{rel}_i)^{1/2}\,]$. Also note that $\tau_i\ =\ \mu_{Ni} + \sigma_{Ni} \cdot (\sqrt{\text{rel}_i})\, \Phi^{-1}[G_i^*(\tau)]$ and $G_i^{-1}[\, G_i^*(\tau) + p] = \mu_{Ni} + \sigma_{Ni}\, \Phi^{-1}[\, G_i^*(\tau) + p]$.

The computation of the y1y2 probability is implemented using the following conditioning argument. For a student having a specified value for $G_1^*(\tau_1)$, condition on a draw of an s_1 from the S_1-distribution $(\, S_1 \mid \tau_1 \ \sim N[\,\tau_1, \sigma_{N1}(1 - \text{rel}_1)^{1/2}\,]\,)$ and express that S_1-value in terms of its fractile of the S_1-distribution, ps_1 , to obtain:

$$\Pr\{\, G_2(S_2) - G_1(S_1) \ \le\ \text{bound} \mid ps_1 \} = \Pr\{\, S_2 \ \le\ G_2^{-1}[G_1(S_1) + \text{bound}] \mid ps_1 \} =$$

$$\Phi\Big[\{\Phi^{-1}[\Phi[(1 - \text{rel}_1)^{1/2}\, \Phi^{-1}[ps_1] + (\sqrt{\text{rel}_1})\, \Phi^{-1}[\, G_1^*(\tau_1)]\,]\, + \text{bound}] - (\sqrt{\text{rel}_2})\, \Phi^{-1}[G_2^*(\tau_2)]\,\}/(1 - \text{rel}_2)^{1/2}\Big] \ . \qquad (2)$$

As a side note to (2), $G_1(S_1)$ can be expressed in terms of ps_1 as:

$$G_1(S_1) = \Phi[(1 - \text{rel}_1)^{1/2}\, \Phi^{-1}[ps_1] + (\sqrt{\text{rel}_1})\, \Phi^{-1}[\, G_1^*(\tau_1)]\,] \ .$$

Then uncondition (2) by integrating $\Pr\{\, G_2(S_2) - G_1(S_1) \le \text{bound} \mid ps_1 \}$ over ps_1 in [0,1]:

$$y1y2 \ =\ \int_0^1 \Big[\Phi\Big[\{\Phi^{-1}[\Phi[(1 - \text{rel}_1)^{1/2}\, \Phi^{-1}[ps_1] + (\sqrt{\text{rel}_1})\, \Phi^{-1}[\, G_1^*(\tau_1)]\,]\, + \text{bound}] - (\sqrt{\text{rel}_2})\, \Phi^{-1}[G_2^*(\tau_2)]\,\}/(1 - \text{rel}_2)^{1/2}\Big]\Big]\, d ps_1 \qquad (3)$$

2.2 Calculations and Illustrations

Maintaining Percentile Rank, $G_1^*(\tau_1) = G_2^*(\tau_2)$. Table 1 displays values of y1y2 in (3) for a student who has maintained percentile rank in year-1 and year-2 in the sense of $G_1^*(\tau_1)$ is set equal to $G_2^*(\tau_2)$. The entries in Table 1 also use the simplification of (3) in setting the year-1 and year-2 test reliability coefficients to be equal, $rel_1 = rel_2$. Consequences of different reliabilities are also discussed below.

Table 1 presents values of y1y2 in (3) for test reliability values from .70 to .95 (.7, .8, .85, .9, .925, .95). For each reliability value the rows of each sub-table represent a student's value of $G_1^*(\tau_1) = G_2^*(\tau_2)$, so that the .60 row indicates a student who "really belongs" at the 60th percentile in both year-1 and year-2. For a test with reliability .90 administered in year-1 and year-2, that student has a 11.6% chance of showing a decrease of at least 20 percentile points and also a 11.6% chance of showing an increase of at least 20 percentile points.

To compare these accuracy results for year-1, year-2 comparisons in Table 1 with more traditional assessments of uncertainty, consider values of the standard error of $G_2(S_2) – G_1(S_1)$. The results for s.e.$[G_2(S_2) – G_1(S_1)]$ are obtained from derivations of the moments of the percentile rank score in Rogosa (1999c). For both year-1 and year-2 tests having reliability .90, s.e.$[G_2(S_2) – G_1(S_1)]$ is 0.1702 for $G_1^*(\tau_1) = G_2^*(\tau_2) = .50$. Increase the test reliabilities to .95 and this s.e.$[G_2(S_2) – G_1(S_1)]$ becomes 0.1231.

Insert Table 1 here

Another version of the statement about the student with $G_1^*(\tau_1) = G_2^*(\tau_2), = .60$ and test reliability .90 is that the probability is .768 that the magnitude of the observed change, $| G_2(S_2) – G_1(S_1) |$, is less than .20. Figure 1 displays these type of probability statements by plotting the values of $\Pr\{ G_2(S_2) – G_1(S_1) \le \text{bound}\} – \Pr\{ G_2(S_2) – G_1(S_1) \le –\text{bound}\}$ as a function of test reliability for $G_1^*(\tau_1) = G_2^*(\tau_2) = .50$ and .75 or .25.

Insert Figure 1 here

The result for y1y2 in (3) allows the year-1 and year-2 test reliability coefficients, rel_1 and rel_2 to differ. The consequences of different test reliabilities can be charted in various ways. Rather than be exhaustive, some specific examples are considered here. With $G_1^*(\tau_1) = G_2^*(\tau_2) = .50$, the effect of differing test reliabilities is minimal in the following set-up. Taking the base as $rel_1 = rel_2 = .90$, the difference in y1y2 values between the base (i.e., Table 1 entries) and differing reliabilities $rel_1 = .85$, $rel_2 = .95$ is less than .001 for the listed values of bound. The same result is found for comparing $rel_1 = rel_2 = .80$ with $rel_1 = .75$, $rel_2 = .85$. Moving away from $G_1^*(\tau_1) = G_2^*(\tau_2) = .50$, differences do emerge. Take $G_1^*(\tau_1) = G_2^*(\tau_2) = .75$, then the

Table 1.
$\Pr\{G_2(S_2) - G_1(S_1) \leq \text{bound} \mid G_1^*(\tau_1),\ G_2^*(\tau_2)\}$ for reliability .70 to .95 and $G_1^*(\tau_1) = G_2^*(\tau_2)$.

Reliability .70

			bound						
	-.20	-.15	-.10	-.05	0.0	.05	.10	.15	.2
$G_1^*(\tau_1)$									
.10	0.124	0.183	0.263	0.369	0.5	0.631	0.737	0.817	0.876
.20	0.186	0.248	0.323	0.408	0.5	0.592	0.677	0.752	0.814
.30	0.219	0.279	0.348	0.422	0.5	0.578	0.652	0.721	0.781
.40	0.235	0.294	0.359	0.428	0.5	0.572	0.641	0.706	0.765
.50	0.24	0.299	0.363	0.43	0.5	0.57	0.637	0.701	0.76
.60	0.235	0.294	0.359	0.428	0.5	0.572	0.641	0.706	0.765
.70	0.219	0.279	0.348	0.422	0.5	0.578	0.652	0.721	0.781
.80	0.186	0.248	0.323	0.408	0.5	0.592	0.677	0.752	0.814
.90	0.124	0.183	0.263	0.369	0.5	0.631	0.737	0.817	0.876

Reliability .80

			bound						
	-.20	-.15	-.10	-.05	0.0	.05	.10	.15	.2
$G_1^*(\tau_1)$									
.10	0.072	0.126	0.211	0.337	0.5	0.663	0.789	0.874	0.928
.20	0.137	0.202	0.287	0.388	0.5	0.612	0.713	0.798	0.863
.30	0.174	0.241	0.319	0.407	0.5	0.593	0.681	0.759	0.826
.40	0.194	0.259	0.333	0.415	0.5	0.585	0.667	0.741	0.806
.50	0.2	0.265	0.338	0.417	0.5	0.583	0.662	0.735	0.8
.60	0.194	0.259	0.333	0.415	0.5	0.585	0.667	0.741	0.806
.70	0.174	0.241	0.319	0.407	0.5	0.593	0.681	0.759	0.826
.80	0.137	0.202	0.287	0.388	0.5	0.612	0.713	0.798	0.863
.90	0.072	0.126	0.211	0.337	0.5	0.663	0.789	0.874	0.928

Reliability .85

			bound						
	-.20	-.15	-.10	-.05	0.0	.05	.10	.15	.2
$G_1^*(\tau_1)$									
.10	0.043	0.089	0.173	0.311	0.5	0.689	0.827	0.911	0.957
.20	0.103	0.168	0.258	0.371	0.5	0.629	0.742	0.832	0.897
.30	0.141	0.21	0.295	0.394	0.5	0.606	0.705	0.79	0.859
.40	0.162	0.23	0.312	0.403	0.5	0.597	0.688	0.77	0.838
.50	0.168	0.237	0.317	0.406	0.5	0.594	0.683	0.763	0.832
.60	0.162	0.23	0.312	0.403	0.5	0.597	0.688	0.77	0.838
.70	0.141	0.21	0.295	0.394	0.5	0.606	0.705	0.79	0.859
.80	0.103	0.168	0.258	0.371	0.5	0.629	0.742	0.832	0.897
.90	0.043	0.089	0.173	0.311	0.5	0.689	0.827	0.911	0.957

Reliability .90

			bound						
	-.20	-.15	-.10	-.05	0.0	.05	.10	.15	.2
$G_1^*(\tau_1)$									
.10	0.017	0.047	0.121	0.27	0.5	0.73	0.879	0.953	0.983
.20	0.06	0.119	0.213	0.344	0.5	0.656	0.787	0.881	0.94
.30	0.096	0.163	0.256	0.372	0.5	0.628	0.744	0.837	0.904
.40	0.116	0.186	0.276	0.383	0.5	0.617	0.724	0.814	0.884
.50	0.123	0.193	0.282	0.387	0.5	0.613	0.718	0.807	0.877
.60	0.116	0.186	0.276	0.383	0.5	0.617	0.724	0.814	0.884
.70	0.096	0.163	0.256	0.372	0.5	0.628	0.744	0.837	0.904
.80	0.06	0.119	0.213	0.344	0.5	0.656	0.787	0.881	0.94
.90	0.017	0.047	0.121	0.27	0.5	0.73	0.879	0.953	0.983

Reliability .925

	bound								
	-.20	-.15	-.10	-.05	0.0	.05	.10	.15	.2
$G_1^*(\tau_1)$									
.10	0.007	0.026	0.087	0.238	0.5	0.762	0.913	0.974	0.993
.20	0.037	0.087	0.18	0.322	0.5	0.678	0.82	0.913	0.963
.30	0.067	0.13	0.226	0.353	0.5	0.647	0.774	0.87	0.933
.40	0.085	0.153	0.248	0.367	0.5	0.633	0.752	0.847	0.915
.50	0.091	0.16	0.254	0.371	0.5	0.629	0.746	0.84	0.909
.60	0.085	0.153	0.248	0.367	0.5	0.633	0.752	0.847	0.915
.70	0.067	0.13	0.226	0.353	0.5	0.647	0.774	0.87	0.933
.80	0.037	0.087	0.18	0.322	0.5	0.678	0.82	0.913	0.963
.90	0.007	0.026	0.087	0.238	0.5	0.762	0.913	0.974	0.993

Reliability .95

	bound								
	-.20	-.15	-.10	-.05	0.0	.05	.10	.15	.2
$G_1^*(\tau_1)$									
.10	0.001	0.008	0.047	0.19	0.5	0.81	0.953	0.992	0.999
.20	0.014	0.048	0.131	0.286	0.5	0.714	0.869	0.952	0.986
.30	0.033	0.085	0.18	0.323	0.5	0.677	0.82	0.915	0.967
.40	0.047	0.106	0.203	0.339	0.5	0.661	0.797	0.894	0.953
.50	0.052	0.113	0.21	0.344	0.5	0.656	0.79	0.887	0.948
.60	0.047	0.106	0.203	0.339	0.5	0.661	0.797	0.894	0.953
.70	0.033	0.085	0.18	0.323	0.5	0.677	0.82	0.915	0.967
.80	0.014	0.048	0.131	0.286	0.5	0.714	0.869	0.952	0.986
.90	0.001	0.008	0.047	0.19	0.5	0.81	0.953	0.992	0.999

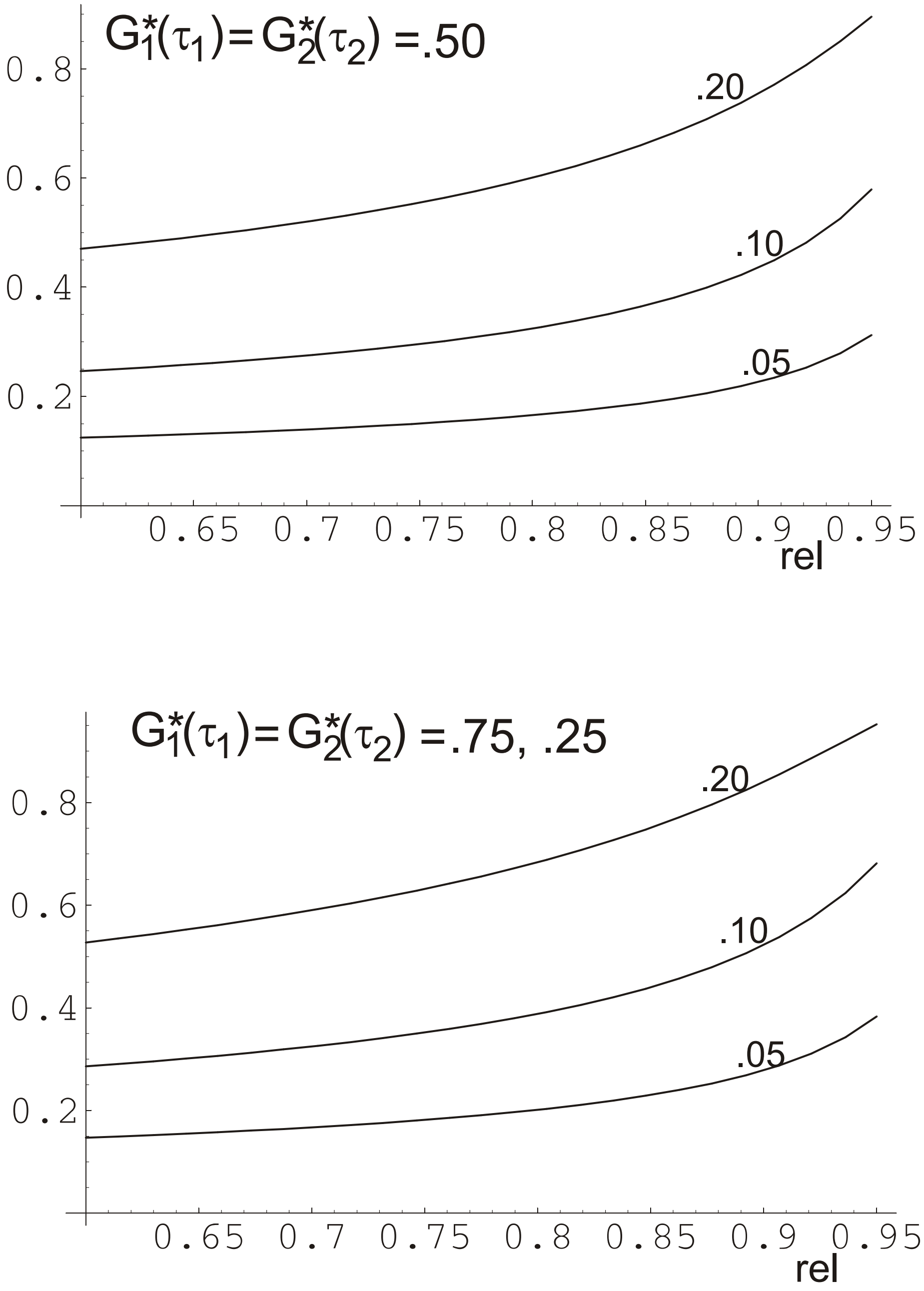

Figure 1. Plots of $\Pr\{ G_2(S_2) - G_1(S_1) \le \text{bound}\} - \Pr\{ G_2(S_2) - G_1(S_1) \le -\text{bound}\}$ as a function of test reliability for $G_1^*(\tau_1) = G_2^*(\tau_2) = .50$ in top frame and $G_1^*(\tau_1) = G_2^*(\tau_2) = .75$ or $.25$ in bottom frame for labeled values of bound $= \{.05, .10, .20\}$.

differences in y1y2 values for $\mathrm{rel}_1 = \mathrm{rel}_2 = .80$ are larger by .02 to .03 than y1y2 values for $\mathrm{rel}_1 = .75$, $\mathrm{rel}_2 = .85$ and smaller by .02 to .03 than y1y2 values for $\mathrm{rel}_1 = .85$, $\mathrm{rel}_2 = .75$. But the two-sided probability statement used in Figure 1, $\Pr\{ G_2(S_2) - G_1(S_1) \le \text{bound}\} - \Pr\{ G_2(S_2) - G_1(S_1) \le -\text{bound}\}$, changes less than .002 for bound = .10 and with $\mathrm{rel}_1 = \mathrm{rel}_2 = .80$ compared to $\mathrm{rel}_1 = .9$, $\mathrm{rel}_2 = .7$ or $\mathrm{rel}_1 = .7$, $\mathrm{rel}_2 = .9$. With bound = .20, the change in y1y2 values is less than .001 for these different reliability configurations. Calculations for year-1, year-2 comparisons based on an actual standardized achievement test, Stanford 9, and in which the tests have different reliabilities (and different norms distributions) can be found in Rogosa (1999b).

Increasing Percentile Rank, $G_1^*(\tau_1) + .10 = G_2^*(\tau_2)$. Table 2 displays values of y1y2 in (3) for a student who improved percentile rank 10 points from year-1 to year-2 in the sense of $G_2^*(\tau_2) = G_1^*(\tau_1) + .10$. The entries in Table 2 also use the simplification of (3) in setting the year-1 and year-2 test reliability coefficients to be equal. One basic question Table 2 informs about is, What's the probability of seeing a decline in the observed percentile rank, even when the student has made a noticeable improvement? (by setting bound = 0 in Table 2). For $G_1^*(\tau_1) = .4$ or .5, that probability is .36 for test reliability .8, .295 for test reliability .9 and .217 for test reliability .95.

Insert Table 2 here

These probability statements, $\Pr\{ G_2(S_2) - G_1(S_1) \le \text{bound} \}$, such as in Table 2 allow a decline in scores to result from $G_1(S_1)$ being "too high" ($G_1(S_1) > G_1^*(\tau_1)$) as much from $G_2(S_2)$ being much "too low." Another view of these kind of calculations can be obtained from Equation (2) by setting $\mathrm{ps}_1 = .5$, which results in a fixing of the value of $G_1(S_1) = G_1(\tau_1) = \Phi[\surd\mathrm{rel}_1)\, \Phi^{-1}[\, G_1^*(\tau_1)]\,]$. Thus $G_2(S_2)$ is the only random component in the student improvement. And for the simplest comparison take $G_1^*(\tau_1) = .50$, as that results in $G_1(S_1) = G_1(\tau_1) = G_1^*(\tau_1) = .50$. By fixing $\mathrm{ps}_1 = .5$ a large random component of $G_2(S_2) - G_1(S_1)$ is removed, and thus we would expect a quantity such as $\Pr\{ G_2(S_2) - G_1(S_1) \le 0\}$ given a "true" increase $G_2^*(\tau_2) = G_1^*(\tau_1) + .10$ would become smaller than the results shown in Table 2 (which are obtained from Equation 3). Below is a comparison of $\Pr\{ G_2(S_2) - G_1(S_1) \le 0\}$ given $G_2^*(\tau_2) = .6$, $G_1^*(\tau_1) = .5$ from Equation (2) (fixing $G_1(S_1) = .5$) and Equation (3). The Equation (2) quantities are smaller, but perhaps not by as much as expected (especially for the lower reliability values).

$\Pr\{ G_2(S_2) - G_1(S_1) \le 0\}$, given $G_2^*(\tau_2) = .6$, $G_1^*(\tau_1) = .5$

```
rel   0.8    0.825  0.85   0.875  0.9    0.925  0.95
eq2   0.306  0.291  0.273  0.251  0.224  0.187  0.135
eq3   0.360  0.349  0.335  0.318  0.295  0.265  0.217
```

Table 2. $\Pr\{ G_2(S_2) - G_1(S_1) \leq \text{bound} \mid G_1^*(\tau_1),\ G_2^*(\tau_2)\}$ for reliability .80 to .95 and $G_1^*(\tau_1) + .10 = G_2^*(\tau_2)$.

Reliability .80

	bound									
	-.20	-.15	-.10	-.05	0.0	.05	.10	.15	.2	.25	.30
$G_1^*(\tau_1)$											
.10	0.032	0.058	0.101	0.169	0.267	0.391	0.522	0.644	0.748	0.829	0.889
.20	0.074	0.114	0.17	0.241	0.327	0.424	0.526	0.624	0.714	0.792	0.855
.30	0.101	0.146	0.203	0.272	0.351	0.437	0.526	0.613	0.696	0.77	0.834
.40	0.114	0.161	0.218	0.285	0.36	0.441	0.525	0.609	0.688	0.76	0.823
.50	0.114	0.161	0.218	0.285	0.36	0.441	0.525	0.609	0.688	0.76	0.823
.60	0.101	0.146	0.203	0.272	0.351	0.437	0.526	0.613	0.696	0.77	0.834
.70	0.074	0.114	0.17	0.241	0.327	0.424	0.526	0.624	0.714	0.792	0.855
.80	0.032	0.058	0.101	0.169	0.267	0.391	0.522	0.644	0.748	0.829	0.889

Reliability .90

	bound										
	-.20	-.15	-.10	-.05	0.0	.05	.10	.15	.2	.25	.30
$G_1^*(\tau_1)$											
.10	0.004	0.011	0.031	0.079	0.175	0.33	0.516	0.69	0.822	0.909	0.958
.20	0.019	0.042	0.083	0.152	0.25	0.377	0.518	0.656	0.774	0.865	0.926
.30	0.035	0.067	0.117	0.188	0.283	0.395	0.518	0.639	0.748	0.837	0.904
.40	0.044	0.079	0.132	0.204	0.295	0.402	0.517	0.632	0.737	0.825	0.892
.50	0.044	0.079	0.132	0.204	0.295	0.402	0.517	0.632	0.737	0.825	0.892
.60	0.035	0.067	0.117	0.188	0.283	0.395	0.518	0.639	0.748	0.837	0.904
.70	0.019	0.042	0.083	0.152	0.25	0.377	0.518	0.656	0.774	0.865	0.926
.80	0.004	0.011	0.031	0.079	0.175	0.33	0.516	0.69	0.822	0.909	0.958

Reliability .95

	bound										
	-.20	-.15	-.10	-.05	0.0	.05	.10	.15	.2	.25	.30
$G_1^*(\tau_1)$											
.10	0.	0.001	0.004	0.021	0.088	0.256	0.511	0.752	0.902	0.97	0.993
.20	0.002	0.007	0.024	0.07	0.164	0.318	0.513	0.703	0.848	0.936	0.978
.30	0.005	0.017	0.045	0.103	0.202	0.343	0.512	0.68	0.817	0.911	0.964
.40	0.008	0.023	0.056	0.118	0.217	0.353	0.512	0.67	0.803	0.899	0.956
.50	0.008	0.023	0.056	0.118	0.217	0.353	0.512	0.67	0.803	0.899	0.956
.60	0.005	0.017	0.045	0.103	0.202	0.343	0.512	0.68	0.817	0.911	0.964
.70	0.002	0.007	0.024	0.07	0.164	0.318	0.513	0.703	0.848	0.936	0.978
.80	0.	0.001	0.004	0.021	0.088	0.256	0.511	0.752	0.902	0.97	0.993

Increasing Percentile Rank, $G_1^*(\tau_1) + .20 = G_2^*(\tau_2)$. Table 3 displays values of y1y2 in (3) for an even larger student improvement from year-1 to year-2 in the sense of $G_2^*(\tau_2) = G_1^*(\tau_1) + .20$. The entries in Table 3 also use the simplification of (3) in setting the year-1 and year-2 test reliability coefficients to be equal. Again, one question to examine is, What's the probability of seeing a decline in the observed percentile rank, even when the student has made a noticeable improvement? Setting bound = 0 in Table 3, for $G_1^*(\tau_1) = .3$, that probability is .229 for test reliability .8, .133 for test reliability .9 and .053 for test reliability .95.

Insert Table 3 here

Table 3. Pr{ $G_2(S_2) - G_1(S_1) \leq$ bound | $G_1^*(\tau_1)$, $G_2^*(\tau_2)$} for reliability .80 to .95 and $G_1^*(\tau_1) + .20 = G_2^*(\tau_2)$.

Reliability .80

	bound										
	-.10	-.05	0.0	.05	.10	.15	.20	.25	.30	.35	.40
$G_1^*(\tau_1)$											
.10	0.05	0.086	0.142	0.221	0.32	0.431	0.543	0.648	0.741	0.818	0.878
.20	0.096	0.142	0.203	0.277	0.362	0.455	0.55	0.642	0.726	0.8	0.861
.30	0.121	0.169	0.229	0.299	0.378	0.463	0.551	0.636	0.717	0.789	0.85
.40	0.128	0.177	0.237	0.306	0.383	0.466	0.551	0.634	0.714	0.785	0.846
.50	0.121	0.169	0.229	0.299	0.378	0.463	0.551	0.636	0.717	0.789	0.85
.60	0.096	0.142	0.203	0.277	0.362	0.455	0.55	0.642	0.726	0.8	0.861
.70	0.05	0.086	0.142	0.221	0.32	0.431	0.543	0.648	0.741	0.818	0.878

Reliability .90

	bound										
	-.10	-.05	0.0	.05	.10	.15	.20	.25	.30	.35	.40
$G_1^*(\tau_1)$											
.10	0.008	0.022	0.054	0.119	0.227	0.371	0.53	0.68	0.802	0.889	0.945
.20	0.028	0.057	0.106	0.18	0.28	0.402	0.535	0.663	0.776	0.864	0.925
.30	0.044	0.079	0.133	0.208	0.303	0.415	0.535	0.654	0.76	0.848	0.912
.40	0.049	0.087	0.141	0.216	0.309	0.418	0.535	0.65	0.755	0.842	0.908
.50	0.044	0.079	0.133	0.208	0.303	0.415	0.535	0.654	0.76	0.848	0.912
.60	0.028	0.057	0.106	0.18	0.28	0.402	0.535	0.663	0.776	0.864	0.925
.70	0.008	0.022	0.054	0.119	0.227	0.371	0.53	0.68	0.802	0.889	0.945

Reliability .95

	bound										
	-.10	-.05	0.0	.05	.10	.15	.20	.25	.30	.35	.40
$G_1^*(\tau_1)$											
.10	0.	0.002	0.01	0.042	0.131	0.3	0.521	0.73	0.876	0.955	0.987
.20	0.003	0.011	0.035	0.089	0.19	0.341	0.524	0.703	0.843	0.931	0.976
.30	0.007	0.021	0.053	0.115	0.217	0.358	0.524	0.688	0.824	0.916	0.967
.40	0.009	0.025	0.059	0.123	0.225	0.363	0.524	0.684	0.817	0.91	0.964
.50	0.007	0.021	0.053	0.115	0.217	0.358	0.524	0.688	0.824	0.916	0.967
.60	0.003	0.011	0.035	0.089	0.19	0.341	0.524	0.703	0.843	0.931	0.976
.70	0.	0.002	0.01	0.042	0.131	0.3	0.521	0.73	0.876	0.955	0.987

Decreasing Percentile Rank, $G_1^*(\tau_1) - .10 = G_2^*(\tau_2)$. Also, the setting in Table 2 can be turned around to examine a student with a "real" decline from year-1 to year-2, in the sense of $G_1^*(\tau_1) > G_2^*(\tau_2)$. Table 4 shows values of y1y2 in (3) for a student with $G_1^*(\tau_1) - .10 = G_2^*(\tau_2)$. The entries in Table 4 also use the simplification of (3) in setting the year-1 and year-2 test reliability coefficients to be equal. Table 4 shows that even with $G_1^*(\tau_1) - .10 = G_2^*(\tau_2)$, the probability of obtaining an increase of 10 or more points in observed percentile rank is a large as .132 for test reliability .90.

Insert Table 4 here

Guaranteeing positive improvement? Not possible, but it is of interest to ask, How much real improvement is needed in order to obtain high probability of an observed improvement? The entries in Table 5 set high probability as .90 and ask how large $k = G_2^*(\tau_2) - G_1^*(\tau_1)$ needs to be in order for $\Pr\{ G_2(S_2) - G_1(S_1) > 0\} = .90$. Even for test reliability .95, k needs to be as large as .165.

Insert Table 5 here

Table 4.
$\Pr\{ G_2(S_2) - G_1(S_1) \leq \text{bound} \mid G_1^*(\tau_1),\ G_2^*(\tau_2)\}$ for reliability .80 to .95 and $G_1^*(\tau_1) - .10 = G_2^*(\tau_2)$.

Reliability .80

	bound								
	-.30	-.25	-.20	-.15	-.10	-.05	0.0	.05	.10
$G_1^*(\tau_1)$									
.20	0.111	0.171	0.252	0.356	0.478	0.609	0.733	0.831	0.899
.30	0.145	0.208	0.286	0.376	0.474	0.576	0.673	0.759	0.83
.40	0.166	0.23	0.304	0.387	0.474	0.563	0.649	0.728	0.797
.50	0.177	0.24	0.312	0.391	0.475	0.559	0.64	0.715	0.782
.60	0.177	0.24	0.312	0.391	0.475	0.559	0.64	0.715	0.782
.70	0.166	0.23	0.304	0.387	0.474	0.563	0.649	0.728	0.797
.80	0.145	0.208	0.286	0.376	0.474	0.576	0.673	0.759	0.83
.90	0.111	0.171	0.252	0.356	0.478	0.609	0.733	0.831	0.899

Reliability .90

	bound								
	-.30	-.25	-.20	-.15	-.10	-.05	0.0	.05	.10
$G_1^*(\tau_1)$									
.20	0.042	0.091	0.178	0.31	0.484	0.67	0.825	0.921	0.969
.30	0.074	0.135	0.226	0.344	0.482	0.623	0.75	0.848	0.917
.40	0.096	0.163	0.252	0.361	0.482	0.605	0.717	0.812	0.883
.50	0.108	0.175	0.263	0.368	0.483	0.598	0.705	0.796	0.868
.60	0.108	0.175	0.263	0.368	0.483	0.598	0.705	0.796	0.868
.70	0.096	0.163	0.252	0.361	0.482	0.605	0.717	0.812	0.883
.80	0.074	0.135	0.226	0.344	0.482	0.623	0.75	0.848	0.917
.90	0.042	0.091	0.178	0.31	0.484	0.67	0.825	0.921	0.969

Reliability .95

	bound								
	-.30	-.25	-.20	-.15	-.10	-.05	0.0	.05	.10
$G_1^*(\tau_1)$									
.20	0.007	0.03	0.098	0.248	0.489	0.744	0.912	0.979	0.996
.30	0.022	0.064	0.152	0.297	0.487	0.682	0.836	0.93	0.976
.40	0.036	0.089	0.183	0.32	0.488	0.657	0.798	0.897	0.955
.50	0.044	0.101	0.197	0.33	0.488	0.647	0.783	0.882	0.944
.60	0.044	0.101	0.197	0.33	0.488	0.647	0.783	0.882	0.944
.70	0.036	0.089	0.183	0.32	0.488	0.657	0.798	0.897	0.955
.80	0.022	0.064	0.152	0.297	0.487	0.682	0.836	0.93	0.976
.90	0.007	0.03	0.098	0.248	0.489	0.744	0.912	0.979	0.996

Table 5. Values of k such that $\Pr\{ G_2(S_2) - G_1(S_1) > 0 \mid G_1^*(\tau_1) + k = G_2^*(\tau_2)\} = .90$ for reliability .80 to .95.

	$G_1^*(\tau_1)$				
	.25	.40	.50	.60	.75
rel					
.80	0.342	0.343	0.318	0.277	0.193
.85	0.285	0.294	0.277	0.245	0.174
.90	0.222	0.237	0.227	0.204	0.149
.95	0.148	0.165	0.161	0.148	0.112

3. Consistency of Percentile Rank Scores Over Years

It depends what the meaning of "consistency" is. Another approach to describing accuracy of the percentile rank scores over successive years is to consider the setting in which $G^*_i(\tau_i)$ is the same over two (or more years), e.g., .60 in year-1 and year-2. And then ask, given constant $G^*_i(\tau_i)$ $(G^*_1(\tau_1) = G^*_2(\tau_2))$, how consistent are the observed $G_1(S_1)$ and $G_2(S_2)$?

In Rogosa (1999a), one approach to describing the accuracy of a percentile rank score was to calculate, for a student whose percentile rank under perfect measurement is $100\, G^*_i(\tau_i)$:

$$\text{hit-rate}_i = \Pr\{\, | G_i(S_i) - G^*_i(\tau_i) | \le \text{tolerance} \,|\, G^*_i(\tau_i)\}.$$

And from Rogosa (1999a), for year *i*

$$\begin{aligned}\text{hit-rate}_i = \; & \Phi[\{\Phi^{-1}[\, G^*_i(\tau_i) + \text{tol}\,] - (\surd\text{rel}_i)\, \Phi^{-1}[\, G^*_i(\tau_i)]\,\}/(1 - \text{rel}_i)^{½}\,] \; - \\ & \Phi[\{\Phi^{-1}[\, G^*_i(\tau_i) - \text{tol}\,] - (\surd\text{rel}_i)\, \Phi^{-1}[G^*_i(\tau_i)]\,\}/(1 - \text{rel}_i)^{½}\,]\ .\end{aligned}$$

And thus a measure of year-to-year consistency is the probability that $G_1(S_1)$ is within the designated closeness to $G^*_1(\tau_1)$ *and* $G_2(S_2)$ is within the designated closeness to $G^*_2(\tau_2)$ (with typically $G^*_1(\tau_1) = G^*_2(\tau_2)$):

$$\text{consistency}_{12} = \text{hit-rate}_1 \cdot \text{hit-rate}_2\ .$$

Table 6 presents values of consistency_{12} for three values of the tolerance: tol = .01, tol=.025, and tol = .05. For example, with a test reliability of .90 for both years, a student "really" at the 60th percentile has probability .026 of both observed percentile rank scores being within 2.5 percentile points of the 60th percentile (i.e., in the range 57.5 to 62.5) and probability .101 of both years' observed percentile rank scores being within 5 percentile points of the 60th percentile.

Insert Table 6 here

Table 6.
Year-1, Year-2 Consistency of Observed Percentile Rank Scores.

tolerance .01

	$G_1^*(\tau_1) = G_2^*(\tau_2)$					
	.25	.40	.50	.60	.75	.90
rel						
0.8	0.003	0.002	0.002	0.002	0.003	0.009
0.825	0.004	0.002	0.002	0.002	0.004	0.011
0.85	0.004	0.003	0.003	0.003	0.004	0.013
0.875	0.005	0.003	0.003	0.003	0.005	0.016
0.9	0.006	0.004	0.004	0.004	0.006	0.02
0.925	0.008	0.006	0.005	0.006	0.008	0.026
0.95	0.012	0.008	0.008	0.008	0.012	0.04

tolerance .025

	$G_1^*(\tau_1) = G_2^*(\tau_2)$					
	.25	.40	.50	.60	.75	.90
rel						
0.8	0.019	0.013	0.012	0.013	0.019	0.058
0.825	0.022	0.015	0.014	0.015	0.022	0.067
0.85	0.026	0.018	0.017	0.018	0.026	0.078
0.875	0.031	0.021	0.02	0.021	0.031	0.094
0.9	0.038	0.026	0.025	0.026	0.038	0.117
0.925	0.051	0.035	0.033	0.035	0.051	0.155
0.95	0.075	0.052	0.049	0.052	0.075	0.224

tolerance .05

	$G_1^*(\tau_1) = G_2^*(\tau_2)$					
	.25	.40	.50	.60	.75	.90
rel						
0.8	0.075	0.052	0.049	0.052	0.075	0.216
0.825	0.085	0.059	0.056	0.059	0.085	0.245
0.85	0.099	0.069	0.065	0.069	0.099	0.281
0.875	0.118	0.082	0.077	0.082	0.118	0.329
0.9	0.145	0.101	0.095	0.101	0.145	0.393
0.925	0.189	0.133	0.125	0.133	0.189	0.486
0.95	0.269	0.192	0.181	0.192	0.269	0.625

References

Assistance Packet for Reporting 1998 STAR Test Results to Parents/Guardians, May 1998, prepared by the Standards, Curriculum, and Assessment Division, California Department of Education.

Assistance Packet for Reporting 1999 STAR Test Results to Parents/Guardians, April 1999, prepared by the Standards, Curriculum, and Assessment Division, California Department of Education.

Rogosa, D.R. (1999a). Accuracy of Individual Scores Expressed in Percentile Ranks: Classical Test Theory Calculations. CRESST Technical Report 509, September, 1999.

Rogosa, D.R. (1999b). How Accurate are the STAR National Percentile Rank Scores for Individual Students?--An Interpretive Guide. July 1999.

Rogosa, D.R. (1999c). Bias and Standard Error of Individual Scores Expressed in Percentile Ranks: Classical Test Theory Calculations. August 1999.