Accuracy of Individual Scores Expressed in Percentile Ranks: Classical Test Theory Calculations

David Rogosa
Stanford University
July 1999

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

Deliverable - July 1999

Project 3.4. Dependability of Assessment Results

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 Individual Scores Expressed in Percentile Ranks: 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 accuracy of the individual percentile rank score. Here, accuracy follows the common-sense interpretation of how close you come to the target. Calculations are presented for: (i) percentile discrepancy, (the difference between the percentile rank of the obtained test score compared to perfectly accurate measurement), (ii) comparisons of a student score to a standard (e.g., national norm 50^{th} percentile), (iii) test-retest consistency (difference between the percentile rank of the obtained test score in two repeated administrations), (iv) comparison of two students (difference between the percentile rank of the obtained test scores for two students of different achievement levels). One important theme is to compare the results of these calculations with the traditional interpretations of the test reliability coefficient: e.g., Does high reliability imply good accuracy?

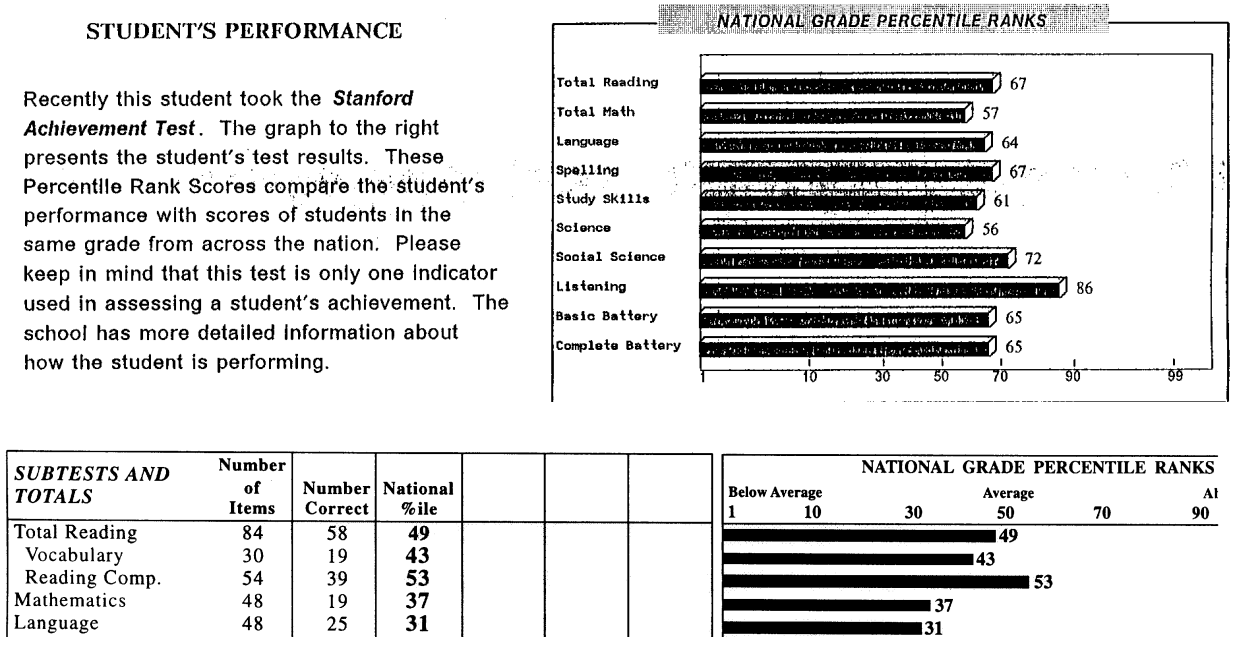
Glossary of Terms and Notation

$G(Y)$	cumulative distribution function of the observed scores Y in the national norming sample
(μ_N , σ_N)	population mean and standard deviation for Y
σ_ε^2	measurement error variance
rel	test reliability coefficient; *rel* = $(\sigma_N^2 - \sigma_\varepsilon^2)/\sigma_N^2$
S,	obtained score for an individual student examinee,
$100 \cdot G(S)$	percentile rank (PR) for the score S
$S = \tau + \varepsilon$	measurement model for S, underlying true score τ
$100 \cdot G(\tau)$	percentile rank in observed norming distribution for individual under perfect measurement
$G_1(Y)$	cumulative distribution function with mean and standard deviation $(\mu_N , [(\sigma_N^2 - \sigma_\varepsilon^2)]^{½})$; $G_1(Y)$ represents a (hypothetical) norming distribution not distorted by measurement error
$100 \cdot G_1(\tau)$	percentile rank for an individual under perfect measurement, in a norming distribution not distorted by measurement error
hit-rate_1	probability that G(S) differs from $G_1(\tau)$ by no more than the tolerance (tol) [section 3.1] $\Pr\{\lvert G(S) - G_1(\tau) \rvert \leq \text{tolerance} \rvert G_1(\tau)\}$
$G^{-1}(P)$	standard set as a percentile, P , of the observed norms distribution. $\Pr\{G(S) > \text{standard}\} = \Pr\{S > G^{-1}(P)\}$ [section 3.2]
retest	test-retest consistency probability, $\Pr\{\lvert G(S_a) - G(S_b) \rvert \leq \text{tolerance} \rvert G_1(\tau)\}$, based on two (contemporaneous) scores for a single student with observed percentile ranks $G(S_a)$ and $G(S_b)$. [section 4]
reversal	reversal probability for two students, $\Pr\{ G(S_1) - G(S_2) > 0 \rvert G_1(\tau_1), G_1(\tau_2)\}$, although $G_1(\tau_1) < G_1(\tau_2)$ [section 5.1]
compare2	probability that the signed difference between the percentile ranks is less than or equal to the quantity "bound" for two students with values of $G_1(\tau_1)$, $G_1(\tau_2)$, $\Pr\{ G(S_1) - G(S_2) \leq \text{bound} \rvert G_1(\tau_1), G_1(\tau_2)\}$. [section 5.2]

1. Introduction

1.1 Reporting Individual Scores as Percentile Ranks

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 California Standardized Testing and Reporting (STAR) program which uses the Stanford Achievement Test Series, Ninth Edition, Form T (Stanford 9), the 1998 *Home Report* and 1999 STAR *Parent Report* feature solely the National Grade Percentile Ranks. (These percentile rank scores also featured in the more extensive *Student Report*, and among the scores reported only the percentile rank is given a graphical display of uncertainty.).



SUBTESTS AND TOTALS	Number of Items	Number Correct	National %ile
Total Reading	84	58	**49**
Vocabulary	30	19	**43**
Reading Comp.	54	39	**53**
Mathematics	48	19	**37**
Language	48	25	**31**

Figure 1. Percentile rank reporting of Stanford 9 results in California STAR; main portion of 1998 STAR *Home Report* at top; excerpt from 1999 *Parent Report* at bottom.

The calculations of this paper seek to provide guidance on the accuracy of test scores reported in the percentile rank metric. How solid are these numbers? Even for tests with respectable raw score reliability coefficients?

1.2 Accuracy in Real Life-- What Is Meant by Accuracy?

The generic statement here is that accuracy follows the common-sense interpretation of how close you come to the target. Television is the main source for these examples of common-sense accuracy. Example 1 is from *Good Housekeeping Institute*, on the accuracy of home body-fat testers, and example 2 is from the Pentagon, on accuracy of cruise missiles. The first example is communicated by Sylvia Chase, ABC News, and the second example by Brian Williams on MSNBC. For the home body-fat testers, the accuracy is expressed in terms of the discrepancy between the home body-fat assessment and the body-fat assessment obtained from much better quality of measurement– a "gold standard" clinical assessment using a body scan. For cruise missiles, the accuracy is stated in terms of the probability that the missile lands "close" (quantified in terms of a distance) to its target.

Home Body-Fat Testers

The first illustration of accuracy is provided by that venerable authority on psychometrics, *Good Housekeeping*. The example is a study of home body-fat testers, conducted by the *Good Housekeeping Institute* reported in the September 1998 issue and also described in the ABC News television program *PrimeTime Live* in August 1998. From the *Good Housekeeping* (p.42) print description:

> Three recent new gizmos promise to calculate your body fat percentage at home. To test their accuracy, Institute chemists sent two female staffers to the weight control unit at St. Luke's-Roosevelt hospital in New York City to have their body fat professionally analyzed. The clinic's results were compared with those of the Tanita body fat monitor and scale, the Omron Body Logic, and the Omron Body Pro.

Good Housekeeping's summative evaluation: "Don't bother, the fat percentages measured by the devices were inconsistent with the clinic's findings." Interestingly, *Good Housekeeping* identified multiple facets of error that influence the home devices: "The amount of fluid in the body, skin temperature, time of day and how long you've been sitting or standing can all interfere with getting an accurate reading".

PrimeTime Live repeated the *Good Housekeeping* tests with an additional 5 volunteers. As in the *Good Housekeeping* trials, the "gold standard Dexa reading" is obtained from the "weight control clinic at New York's St. Luke's Roosevelt Hospital, [with] the Rolls Royce of body fat analyzers -- the Dexa, considered the most accurate of fat measuring devices.... The Dexa scans the body, sorting out what's fat and what's not."(Primetime Live 8/12/98). For one female subject the Dexa gave 33 percent body fat. However, the Omron gave a 24 percent reading and the health club skin-fold with calipers also gave 24 percent (recommended upper limit is 25 percent). For one male subject, the Dexa gave 15.9 percent, whereas skin-fold gave 5 percent.

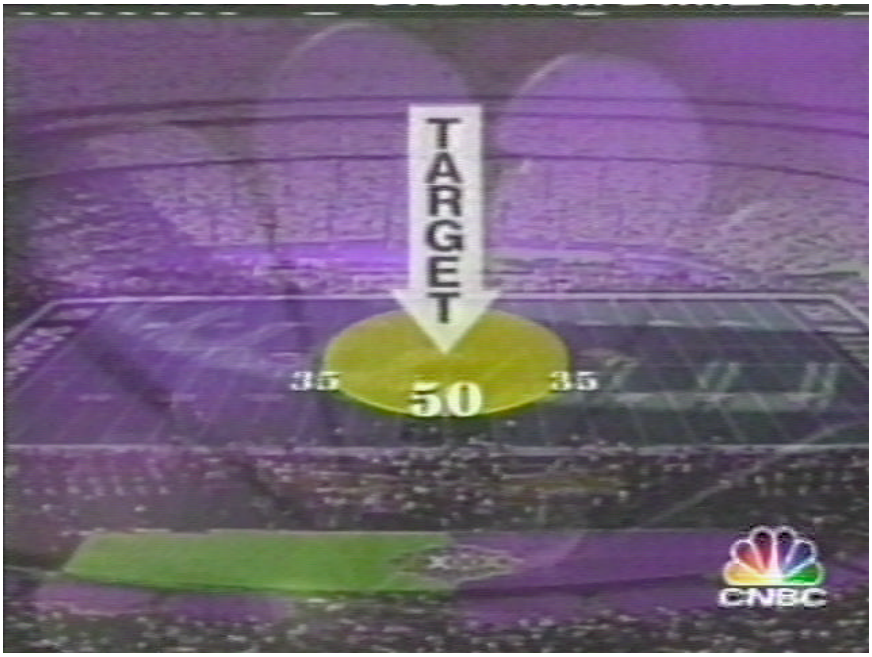
The intended lesson from the body fat example is that the natural way to evaluate accuracy of a measurement, whether it be a percentile rank score from a standardized test or a reading from a home body fat tester, is by the discrepancy between the gold-standard assessment (here the Dexa reading) and the field reading (here the home device). In the *Good Housekeeping* trials, the approach is if home tester produces scores close to clinical body fat evaluation, then it's a good buy. Whether the observed discrepancies are acceptably small is a matter for judgement; in these trials it seems a discrepancy of 10 percent body fat is viewed as much too large to recommend the home devices or skin-fold. Extending this example, envision a far more extensive evaluation of the home body fat testers, in which, say, 1,000 individuals had a gold-standard reading from the Dexa and also measurements from each of the home devices. From those hypothetical data, the proportion of measurements within 5 percentage points of the Dexa, within 10 percentage points of the Dexa, etc., for each device, could be tabulated. That's the type of assessment (via probability calculations) that will be presented in the next section for the standardized test percentile rank score.

Cruise Missile Accuracy

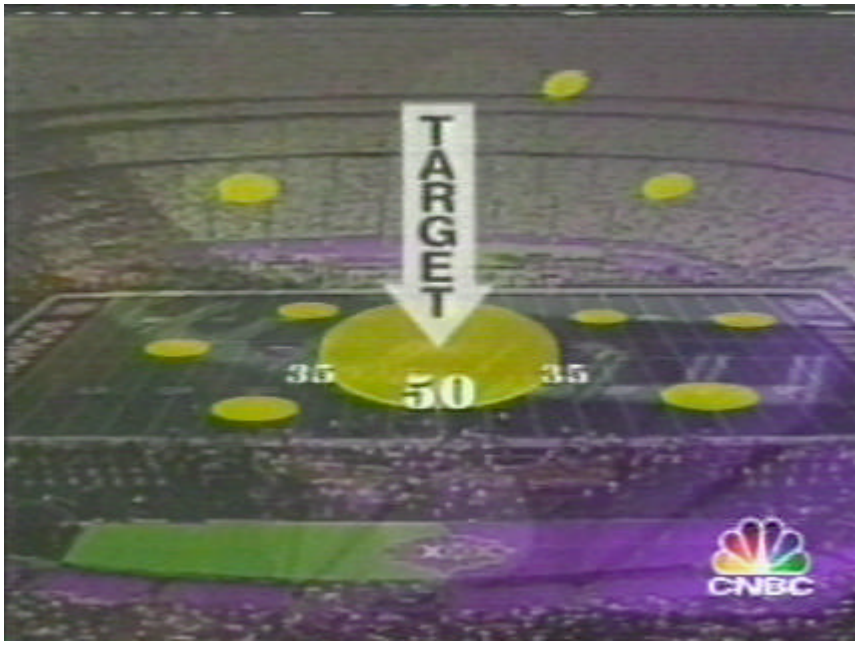
The second illustration of accuracy is provided by descriptions of the accuracy of the Tomahawk cruise missile in a November 12, 1998, segment of the MSNBC program *News with Brian Williams*, titled Tomahawk Diplomacy. The screenshots in Figure 2 are accompanied by the narration at the right.

Insert Figure 2 here

To recast in the terms we will use for accuracy of percentile rank test scores, the top frame of Figure 2 indicates the hit-rate is .50 for target at the 50-yard-line, and tolerance 15 yards. In military jargon, the acronym is CEP, which stands for Circular Error Probable-- a measure of the radius of the circle into which 50 percent of weapons should impact. The bottom frame isn't exactly quantifiable in terms of hit rate, but roughly we could say: hit-rate is large (e.g.,$\geq$ .9) for strike within the confines of the playing field and hit-rate is very large (e.g.,$\geq$ .98) for strike within the stadium. (A narrative version of this description of Tomahawk cruise missile accuracy is provided, for example, by a DOD News Briefing by K. Bacon, 9/6/96.)



"The Pentagon uses the idea of a football field. Says if the target were the 50 yard line, half the missiles would hit within fifteen yards



most of the rest [fall] on the field, but a few in the stands or even outside the stadium"

(MSNBC, Nov. 12, 1998).

Figure 2. Screenshots and accompanying narration on the accuracy of the Tomahawk cruise missile in a November 12, 1998, segment of the MSNBC program *News with Brian Williams*, titled Tomahawk Diplomacy.

The analogy that is used here for the accuracy of percentile rank scores is, What's the probability that the obtained percentile rank score lies within 5 percentile points of the target?, or 10 percentile points of the target? Definition of the target for the percentile rank score is through the (hypothetical) gold standard measurement obtained from a far more extensive testing protocol (or repeated testings) of student achievement. The technical content of this paper addresses the question: For the standardized testing situation, will the percentile rank scores have adequate accuracy? Or to refine the question, for what levels of the reliability coefficient will adequate accuracy be obtained? One way to display the information on accuracy is through the probability that the discrepancy between the percentile rank of the obtained student score and that for the gold standard measurement is less than a specified tolerance (see Section 3.1).

2. Technical Formulation

The technical formulation is a basic errors-in-variables model with all components having Gaussian distributions. All that is meant by the phrase "classical test theory calculation" is to identify the calculations herein as pertaining to the simplest case of constant error variance across the score distribution with continuous, Normally distributed scores. Thus, this formulation represents a restricted version of the full statement of the classical test theory model, see Lord (1980, Chap. 1). The components of what is referred to as the classical test theory calculation are listed below.

- The cumulative distribution function of the observed scores Y in the national norming sample is denoted by G(Y). The classical test theory formulation defines this norming distribution, with density function g(Y), to be a Normal Distribution; denote the corresponding population mean and standard deviation for Y by (μ_N , σ_N).

- 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 σ_ε^2 across the score distribution: i.e., $\varepsilon \sim N(0, \sqrt{\sigma_\varepsilon^2})$. (More general formulations, such as σ_ε^2 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 $rel = (\sigma_N^2 - \sigma_\varepsilon^2)/\sigma_N^2$. 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 approximate 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 distribution, G(Y) is 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 $G_1(Y)$ the cumulative distribution function with corresponding mean and standard deviation $(\mu_N , [(\sigma_N^2 - \sigma_\varepsilon^2)]^{1/2})$; $G_1(Y)$ 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 $100 \cdot G(S)$; thus G(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 $100 \cdot G(\tau)$ or, in a norming distribution not distorted by measurement error, the percentile rank is $100 \cdot G_1(\tau)$. Often in the calculations, an individual (or an individual's achievement level) is characterized a value of $G_1(\tau)$.

Adopting the interpretation of accuracy as how close you come to the target, the following sections present calculations examining the accuracy of the individual percentile rank score by comparing obtained score 100 G(S) to the percentile rank score that would be obtained with perfectly accurate measurement 100 $G_1(\tau)$ or 100 $G(\tau)$ (depending whether the calculation is done for a norming distribution not distorted by measurement error). Section 3 focuses on the difference between G(S) and $G_1(\tau)$, for $G_1(\tau)$ set to a specific percentile: e.g., values of G(S) for a student with true standing at 50^{th} percentile as in $\Pr\{|G(S) - .50| \leq .10\}$. Also, Section 3 includes calculations comparing a student to a standard, such as a standard set at observed 50^{th} percentile; calculate, for example, probability above standard for student with true percentile rank, $G_1(\tau) = .60$. Section 4 extends the calculations to test-retest consistency; for example, for a student with $G_1(\tau) = .50$ (true standing at 50^{th} percentile), obtain two test scores (S_a and S_b) and calculate $\Pr\{|G(S_a) - G(S_b)| < .10\}$. In Section 5, those calculations are extended to comparing observed percentile scores from two students at different achievement levels. One scenario, illustrated in Figure 4, has student 1 at 50^{th} percentile ($G_1(\tau_1) = .50$) and student 2 at 75^{th} percentile ($G_1(\tau_2) = .75$) (under perfect measurement), and, for example, calculate the probability of a reversal, $\Pr\{G(S_1) - G(S_2) > 0\}$.

Figures 3 and 4 provide a depiction of the components of the calculations. Figure 3 shows the formulation relevant to Sections 3 and 4, whereas Figure 4 is relevant to Section 5. A visual harbinger of the results in Sections 3 and 4 is provided by noting in Figure 3, even with test reliability .90, that the depicted score distribution conditional on the true percentile (pdf of S) does span a wide range of values of G(Y).

Insert Figures 3 and 4 here

One important theme is to compare the results of these calculations with the traditional interpretations of the test reliability coefficient: e.g., Does high reliability imply good accuracy? One way to address this question is through results of the calculations described above for test reliability in the range of .80 to .95, which represent values for the best standardized tests and subtests.

The present calculations are deliberately reduced to the most abstracted situation for ease of exposition, and because the main points about accuracy are adequately illustrated by the abstracted formulation. Similar calculations for specific (e.g., IRT scaled) tests with error variances differing across the score distribution and empirically obtained norms distributions (not necessarily Gaussian) and with discrete scale score points can be carried out with some added complexity.

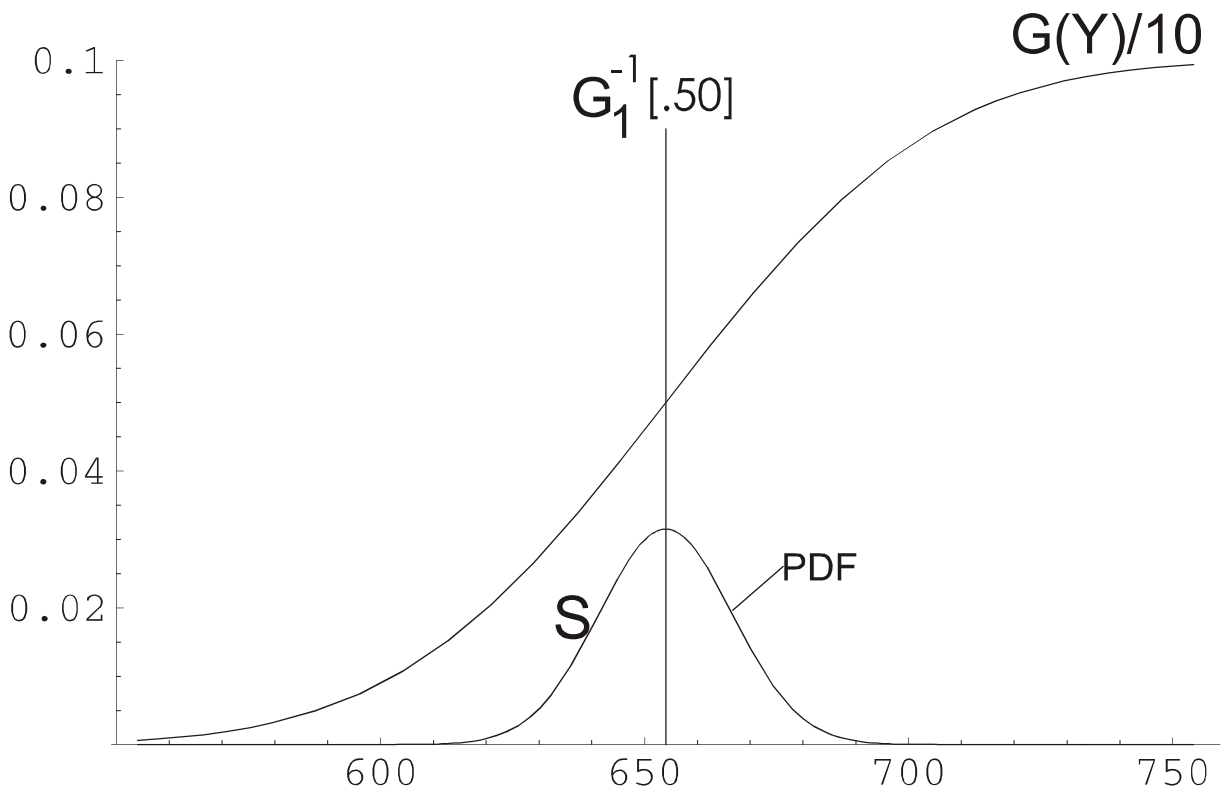


Figure 3. Diagram for calculations for percentile accuracy for observed student score (S) for student with true standing at 50th percentile. Diagram uses test reliability .90 and true score distribution N(654,38). Diagram also represents test-retest consistency calculations with two draws from S distribution.

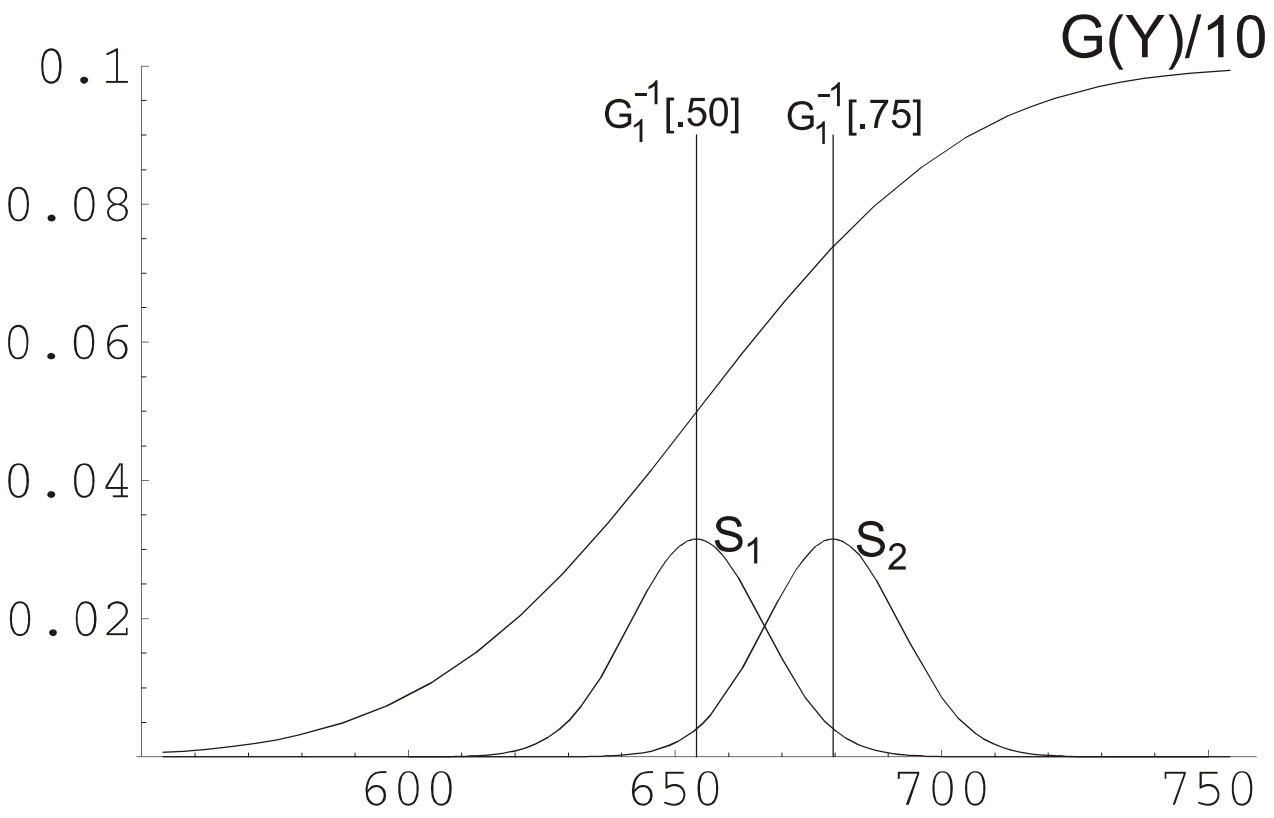


Figure 4. Diagram for calculations for comparing scores for student 1 (score distribution S_1) at 50^{th} percentile and student 2 (score distribution S_2) at 75^{th} percentile (under perfect measurement). Diagram uses test reliability .90 and true score distribution N(654,38).

3. Accuracy for a Single Student Score

3.1 Percentile Discrepancy

How solid is the observed student percentile rank score 100·G(S)? Accuracy of the percentile rank is assessed by calculations based on the discrepancy between G(S) and the percentile rank score the student would receive if measurement were without error. Taking the target to be the percentile rank of the true student score for a norming distribution not distorted by measurement error, the hit-rate is defined in terms of $G(S) - G_1(\tau)$. That is, calculate for a student whose percentile rank under perfect measurement is $100 \cdot G_1(\tau)$:

$$\text{hit-rate}_1 = \Pr\{\,|\, G(S) - G_1(\tau) \,|\, \leq \text{tolerance} \,|\, G_1(\tau)\}. \qquad (3.1)$$

Hit-rate_1 is the probability that G(S) differs from $G_1(\tau)$ by no more than the tolerance; this probability is put forth as one of the major ways to assess the accuracy of student percentile rank score $100 \cdot G(S)$. To calculate hit-rate_1 for a test with reliability *rel*, specify the student's percentile rank under perfect measurement (e.g., $G_1(\tau) = .50$ or .35) and specify the tolerance, *tol* (e.g., .10). Figure 3 provides a depiction for the core calculation with $G_1(\tau) = .5$. Extended numerical illustrations follow the derivation below in a series of Exhibits (e.g. $G_1(\tau) = .5$ in Exhibit IA and $G_1(\tau) = .75, .25$ in Exhibit IB).

Derivation.

$$\text{hit-rate}_1 = \Pr\{\, G^{-1}[G_1(\tau) - \text{tol}] \leq S \leq G^{-1}[G_1(\tau) + \text{tol}]\,\}$$

$$= \Pr\{\, S \leq G^{-1}[G_1(\tau) + \text{tol}]\,\} - \Pr\{\, S < G^{-1}[G_1(\tau) - \text{tol}]\,\}$$

Let $\Phi[x]$ indicate the distribution function (cdf) for N(0,1) and $\phi[x]$ indicate the density (pdf) for N(0,1). Then $G(x) = \Phi[(x - \mu_N)/\sigma_N]$, and $S \,|\, \tau \sim N[\,\tau\,, \sigma_N(1 - \text{rel})^{1/2}\,]$ so that $\Pr\{S \leq x\} = \Phi[(x - \tau)/\,\sigma_N(1 - \text{rel})^{1/2}\,]$. Also note that $\tau = \mu_N + \sigma_N (\sqrt{\text{rel}})\, \Phi^{-1}[G_1(\tau)]$ and $G^{-1}[G_1(\tau) + \text{tol}] = \mu_N + \sigma_N \Phi^{-1}[G_1(\tau) + \text{tol}]$.

Then substituting into hit-rate_1 yields

$$\text{hit-rate}_1 = \Phi[\{\Phi^{-1}[\, G_1(\tau) + \text{tol}\,] - (\sqrt{\text{rel}})\, \Phi^{-1}[G_1(\tau)]\,\}/(1 - \text{rel})^{1/2}] - \Phi[\{\Phi^{-1}[\, G_1(\tau) - \text{tol}\,] - (\sqrt{\text{rel}})\, \Phi^{-1}[G_1(\tau)]\,\}/(1 - \text{rel})^{1/2}] \qquad (3.2)$$

Equation 3.2 shows that hit-rate_1 depends only on rel, the test reliability coefficient, $G_1(\tau)$, the level of the student true measurement in the norming distribution not distorted by measurement error, and, tol, the chosen tolerance. In numerical computations, if $G_1(\tau) + \text{tol} > 1$ set it to 1, and if $G_1(\tau) - \text{tol} < 0$, set it to 0.

Alternatively, defining the target to be $G(\tau)$, recast the hit-rate in terms of the distance between G(S) and $G(\tau)$ (i.e. not considering the distortion of the norming distribution by the error of measurement). We can write $G(\tau) = \Phi[\ (\sqrt{rel})\ \Phi^{-1}[G_1(\tau)]\ ]$ or $G_1(\tau) = \Phi\ [\Phi^{-1}[G(\tau)]/\sqrt{rel}\ ]$ to obtain:

$$\begin{aligned} \text{hit-rate}_G &= \Pr\{|\,G(S) - G(\tau)\,| \le \text{tolerance}\} \\ &= \Phi[\{\Phi^{-1}[\,G(\tau) + \text{tol}\,] - \Phi^{-1}[G(\tau)]\,\}/(1-\text{rel})^{1/2}] - \\ &\quad \Phi[\{\Phi^{-1}[\,G(\tau) - \text{tol}\,] - \Phi^{-1}[G(\tau)]\,\}/(1-\text{rel})^{1/2}] \end{aligned} \qquad (3.3)$$

With $G_1(\tau) = .5$, hit-rate$_1$ and hit-rate$_G$ are identical. Otherwise the difference between (3.2) and (3.3), for the same τ, is small-to-negligible in most situations, except for low reliability and $G_1(\tau)$ near 0 or 1. So it appears sufficient to illustrate percentile discrepancy using only values for hit-rate$_1$.

Exhibits IA-IE presents results, expressed in Tables and Figures, for the percentile discrepancy in terms of hit-rate$_1$. The exposition here attempts to give a guided tour of the displayed results in the Exhibits, and encourages readers to draw their own conclusions about acceptable levels of accuracy.

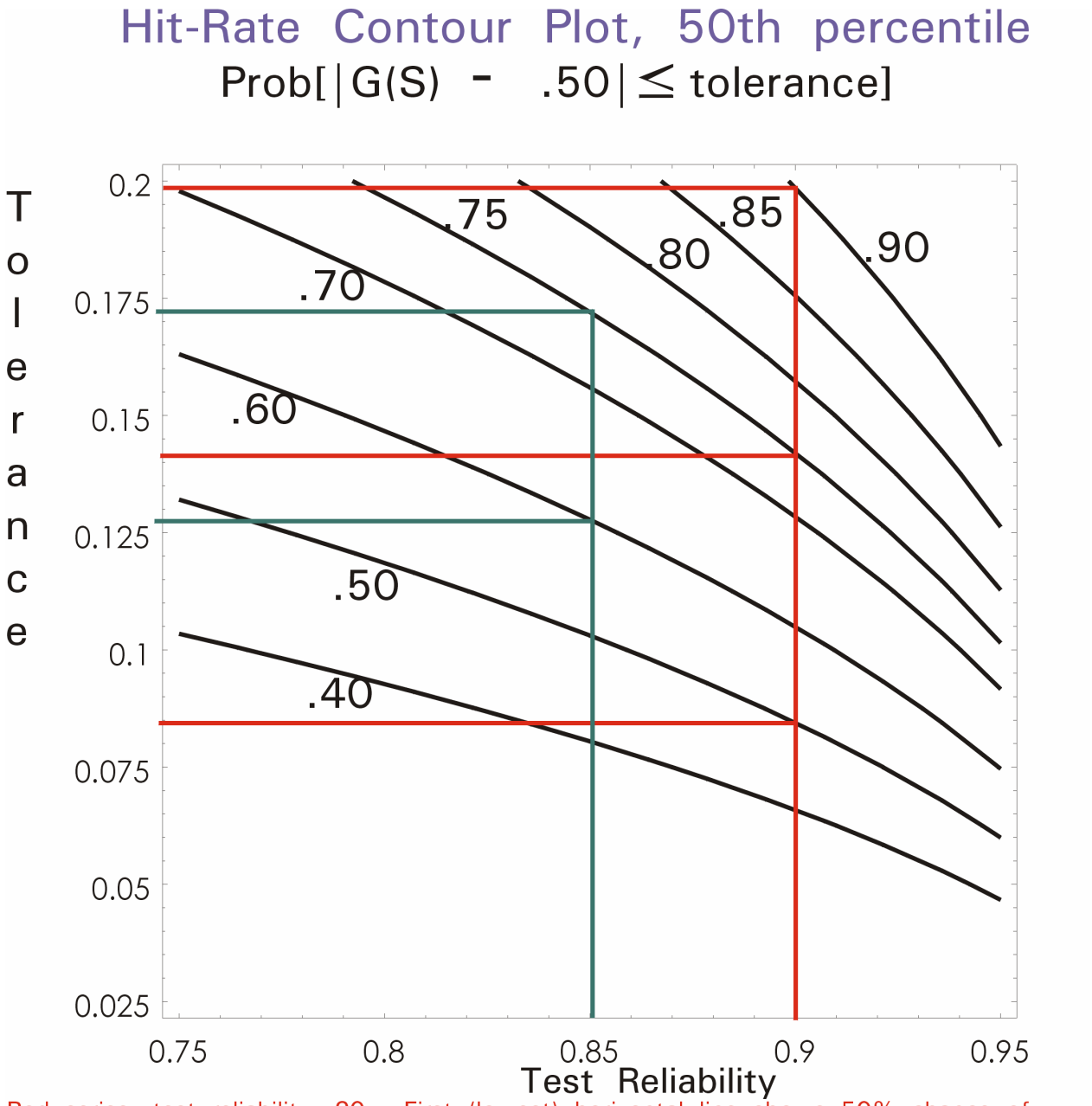
Percentile .50. For a student with true standing at 50th percentile ($G_1(\tau) = .50$), hit-rate$_1$ is $\Pr\{|\,G(S) - .50\,| \le \text{tolerance}\}$. Exhibit IA presents various forms of that calculation, which depends on the test reliability and the specified value of tolerance. The 3D plot in Figure IA1 gives a broad look at the dependence of the hit-rate on test reliability and the value of tolerance. The entries in Table IA1 in Exhibit 1A give the numerical values for the 3D plot. For test reliability .90, the hit-rate is .309 for tolerance .05, .577 for tolerance .10, and .777 for tolerance .15. That is, the probability that the observed percentile rank is within 10 percentile points of $G_1(\tau) = .5$ (i.e., G(S) between .40 and .60) is .577 for test reliability .90 (and hit-rate is reduced .487 for test reliability .85). Does that seem to be acceptable accuracy? What reliability seems necessary to obtain acceptable accuracy? Increasing the reliability to .95 (which is equivalent to more than doubling the test length) increases the hit-rate noticeably (but maybe not dramatically): For tolerance .05, hit-rate increases from .309 to .426; for tolerance .10, hit-rate increases from .577 to .743

A more traditional way to approach the accuracy of G(S) is to think in terms of possible bias in and the magnitude of the standard error of G(S). In a separate report, I derive exact expressions for the moments of G(S); for purposes here it may be useful to report that the standard error of G(S) for a student with $G_1(\tau) = .50$ and with test reliability .90 is .1204 (twelve percentile points). This result can almost be read off of Figure 3. Even for test reliability .95, the standard error of G(S) is .0871, and for test reliability .85 the standard error of G(S) is .1443.

The dependence of hit-rate on test reliability is most directly shown in Figure IA2; each plotted curve is labeled with a specified level for the tolerance. The final entry in Exhibit IA is the contour plot in Figure IA3; each of the contours is labeled with a specified level for the hit-rate. The purpose of the contour plot is to show the combination of test reliability and tolerance needed to achieve a specific value for the hit-rate. A guide to reading the contour plot is provided by Figure 5.

One additional version of these accuracy calculations follows from the CEP (circular error probable) from the cruise missile example. To obtain hit-rate$_1$ = .50 with $G_1(\tau)$ = .50 requires tolerance .144 for test reliability .7, tolerance .119 for test reliability .8, tolerance .084 for test reliability .9, and tolerance .06 for test reliability .95. But is hit-rate$_1$ = .50 an adequate enough accuracy standard for judging the performance of individual schoolchildren?

Insert Exhibit IA here
Insert Figure 5 here



Red series, test reliability .90. First (lowest) horizontal line shows 50% chance of observed percentile rank score being within ± 8.5% of true percentile.
Next (middle) horizontal line shows 75% chance of observed percentile rank score being within ± 14% of true percentile (e.g. Width nearly 28%, more than a whole quartile).
Top horizontal line shows 90% chance of observed percentile rank score being within ± 19.5% of true percentile (e.g. Width 40%).
Green series, test reliability .85. Bottom horizontal line shows 60% chance of observed percentile rank score being within ± 13% of true percentile (e.g. Width 25%, a whole quartile).
Top horizontal line shows 75% chance of observed percentile rank score being within ± 17% of true percentile (e.g. Width 34%).

Figure 5. Color-annotated version of the contour plot in Figure IA3.

Exhibit IA
Percentile Accuracy for student at 50th percentile

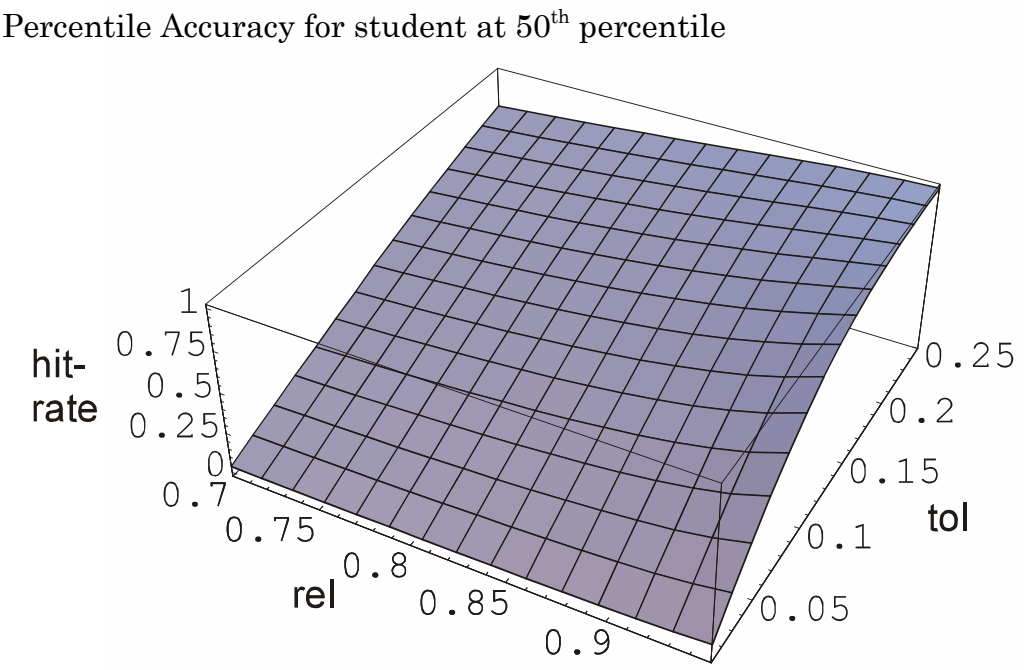


Figure IA1. 3D plot of Prob within tolerance (hit-rate) as a function of reliability (.7, .95), and tolerance (.025, .25) for student at 50th percentile.

Table IA1. Hit-rate as a function of reliability (.7, .95), and tolerance (.025, .25) for student at 50th percentile.

	tolerance				
	.025	.05	.075	.10	.125
rel					
0.70	0.0911	0.181	0.27	0.356	0.439
0.725	0.0952	0.189	0.282	0.371	0.457
0.75	0.0998	0.198	0.295	0.388	0.476
0.775	0.105	0.209	0.31	0.407	0.498
0.8	0.112	0.221	0.328	0.429	0.524
0.825	0.119	0.236	0.349	0.455	0.554
0.85	0.129	0.254	0.375	0.487	0.589
0.875	0.141	0.278	0.407	0.526	0.633
0.9	0.157	0.309	0.45	0.577	0.686
0.925	0.181	0.354	0.51	0.645	0.755
0.95	0.221	0.426	0.602	0.743	0.846

	tolerance				
	.15	.175	.20	.225	.25
rel					
0.70	0.518	0.593	0.662	0.725	0.782
0.725	0.538	0.613	0.683	0.746	0.802
0.75	0.559	0.636	0.706	0.768	0.823
0.775	0.583	0.661	0.731	0.792	0.845
0.8	0.611	0.69	0.759	0.819	0.868
0.825	0.643	0.722	0.79	0.847	0.893
0.85	0.68	0.759	0.824	0.877	0.918
0.875	0.724	0.801	0.862	0.909	0.944
0.9	0.777	0.849	0.903	0.941	0.967
0.925	0.841	0.902	0.944	0.971	0.986
0.95	0.915	0.958	0.981	0.992	0.997

Exhibit IA continued

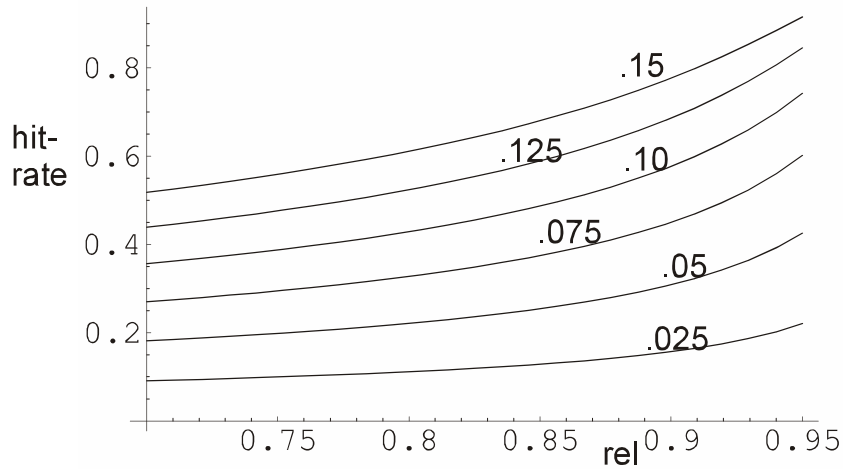
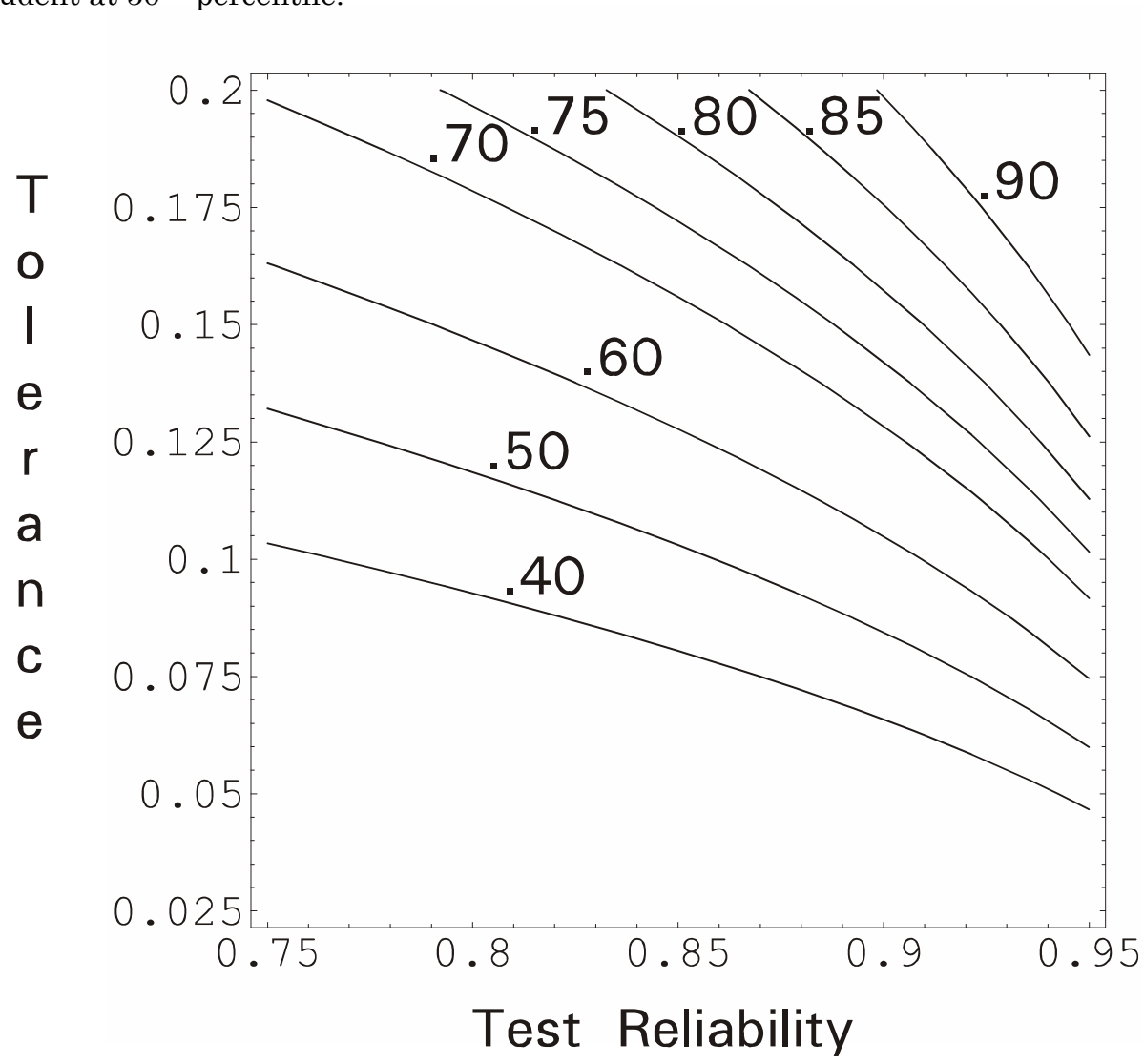


Figure IA2. Plot of hit-rate as a function of reliability (.7, .95), at each level of tolerance (.025, .05, .075, .10, .125, .15) for student at 50th percentile.

Figure IA3. Contour plot of hit-rate as a function of tolerance and test reliability for student at 50th percentile.



Percentiles 75, 25. Because the 50th percentile ($G_1(\tau) = .50$) produces the worst results for hit-rate, Exhibit IB repeats the full set of displays for the 75th and 25th percentiles; hit-rate results are identical for $G_1(\tau) = .75$ and $G_1(\tau) = .25$. Hit-rates are higher for the same test reliability and tolerance in Exhibit IB than in Exhibit IA even though the measurement error variance is the same, because the percentile rank scores depend on G(Y) (which flattens out for higher and lower percentiles). For test reliability .90 the hit-rate moves up to .381 (from .309) for tolerance .05, and to .685 (from .577) for tolerance .10. Taking the ratio of the entries of Table IB1 to those of Table IA1, the hit-rate for $G_1(\tau) = .75, .25$ is about 1.2 times as large as that for $G_1(\tau) = .50$ for the lower values of tolerance, the ratio decreasing to about 1.05 for high reliability and larger tolerance values. For the CEP calculations, to obtain $\text{hit-rate}_1 = .50$ with $G_1(\tau) = .75$ or .25 requires tolerance .117 for test reliability .7, tolerance .095 for test reliability .8, tolerance .068 for test reliability .9, and tolerance .048 for test reliability .95.

Insert Exhibit IB here

Other $G_1(\tau)$ *values.* The basic results in Exhibits IA and IB can be augmented by additional displays for other choices of $G_1(\tau)$; i.e., How much do the results in Exhibit IA change for other choices of $G_1(\tau)$? Exhibit IC gives plots of the hit-rate with each curve labeled for percentiles 50, 60, 70, 80, 90. (Only curves for $G_1(\tau) \geq .50$ are used for the comparison, as results for $G_1(\tau) = .2$ would be identical to $G_1(\tau) = .8$ and so forth.) First is the set of plots for test reliability .90, .85, .80 in Figure IC1. Those plots of hit-rate vs tolerance show 50th and 60th percentiles are almost indistinguishable; $G_1(\tau) = .80$ gives notably higher hit-rates than $G_1(\tau) = .60$. Next is a similar set of plots for hit-rate vs test reliability for chosen levels of tolerance in Figure IC2. Furthermore, Exhibit ID gives the table of hit-rate for test reliability and tolerance shown in Table IA1 for each of percentiles 60, 70, 80, 90. Next is the plot in Figure ID1 of hit-rate on test reliability for $G_1(\tau) = .90$ (compare with $G_1(\tau) = .50$ in Figure IA2). Finally, Figure ID2 displays the set of contour plots (shown for $G_1(\tau) = .50$ in Figure IA3) for each of $G_1(\tau) = .60, .70, .80, .90$.

Insert Exhibits IC, ID here

Randomly-sampled student. The final exhibit on these hit-rate calculations is constructed for a student sampled at random from the population of students; for example, consider drawing a $G_1(\tau)$ at random from U[0,1]. The sampling could be described as Percentile Accuracy averaged over all percentiles. In Exhibit IE, the results for hit-rate are slightly higher than results for $G_1(\tau) = .75, .25$ in Exhibit IB.

Insert Exhibit IE here

Exhibit IB
Percentile Accuracy for student at 75th percentile

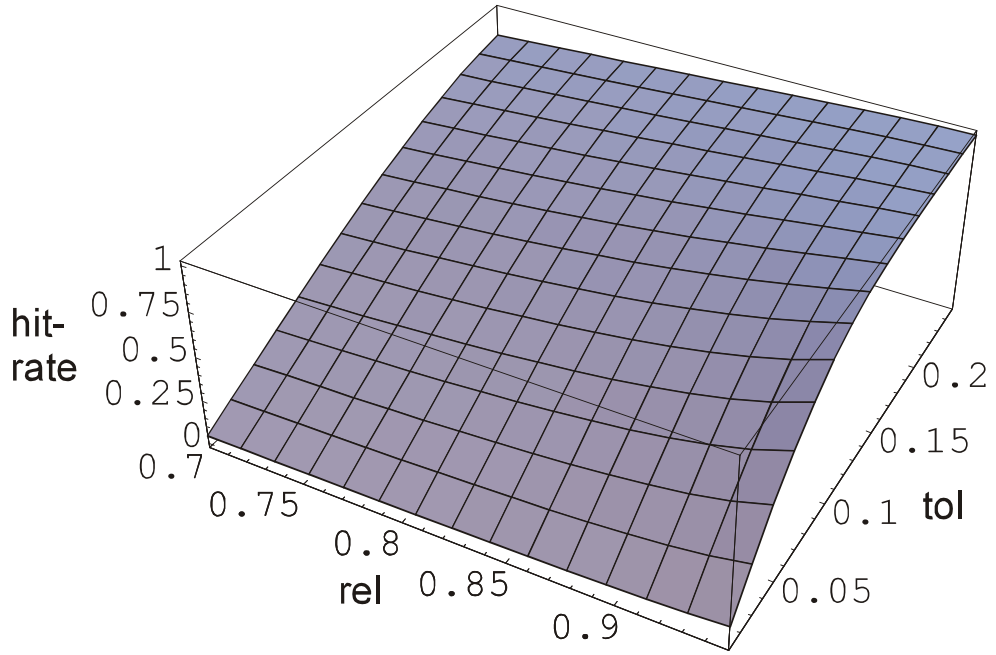


Figure IB1. Hit-rate as a function of reliability (.7, .95), and tolerance (.025, .25) for student at 75th percentile.

Table IB1: Hit-rate as a function of reliability (.7, .95), and tolerance (.025, .25) for student at 75th percentile.

	tolerance				
	.025	.05	.075	.10	.125
rel					
0.70	0.112	0.223	0.33	0.434	0.531
0.725	0.117	0.233	0.345	0.452	0.551
0.75	0.123	0.244	0.361	0.472	0.574
0.775	0.13	0.257	0.38	0.494	0.599
0.8	0.138	0.273	0.402	0.521	0.627
0.825	0.148	0.291	0.427	0.551	0.66
0.85	0.16	0.314	0.458	0.587	0.697
0.875	0.175	0.343	0.497	0.631	0.741
0.9	0.196	0.381	0.546	0.685	0.792
0.925	0.225	0.434	0.613	0.753	0.852
0.95	0.275	0.519	0.711	0.843	0.921

	tolerance				
	.15	.175	.20	.225	.25
rel					
0.70	0.62	0.698	0.764	0.815	0.849
0.725	0.641	0.719	0.783	0.831	0.863
0.75	0.664	0.742	0.803	0.849	0.879
0.775	0.69	0.766	0.825	0.867	0.895
0.8	0.718	0.792	0.847	0.885	0.911
0.825	0.75	0.82	0.871	0.905	0.928
0.85	0.785	0.851	0.896	0.925	0.946
0.875	0.824	0.883	0.921	0.946	0.963
0.9	0.868	0.917	0.947	0.966	0.978
0.925	0.915	0.951	0.972	0.984	0.991
0.95	0.962	0.982	0.991	0.996	0.998

Exhibit IB continued

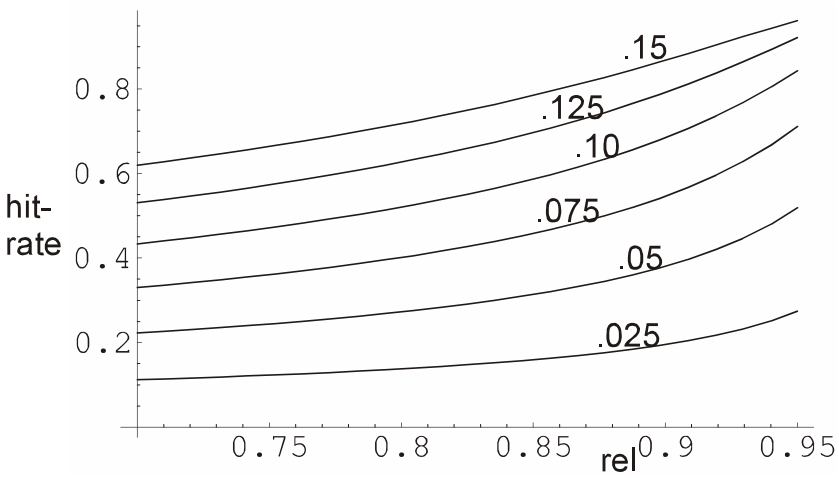


Figure IB2. Plot of hit-rate as a function of reliability (.7, .95), at each level of tolerance (.025, .05, .075, .10, .125, .15) for student at 75th percentile.

Figure IB3. Contour plot of hit-rate as a function of tolerance and test reliability for Student at 75th , 25th percentile.

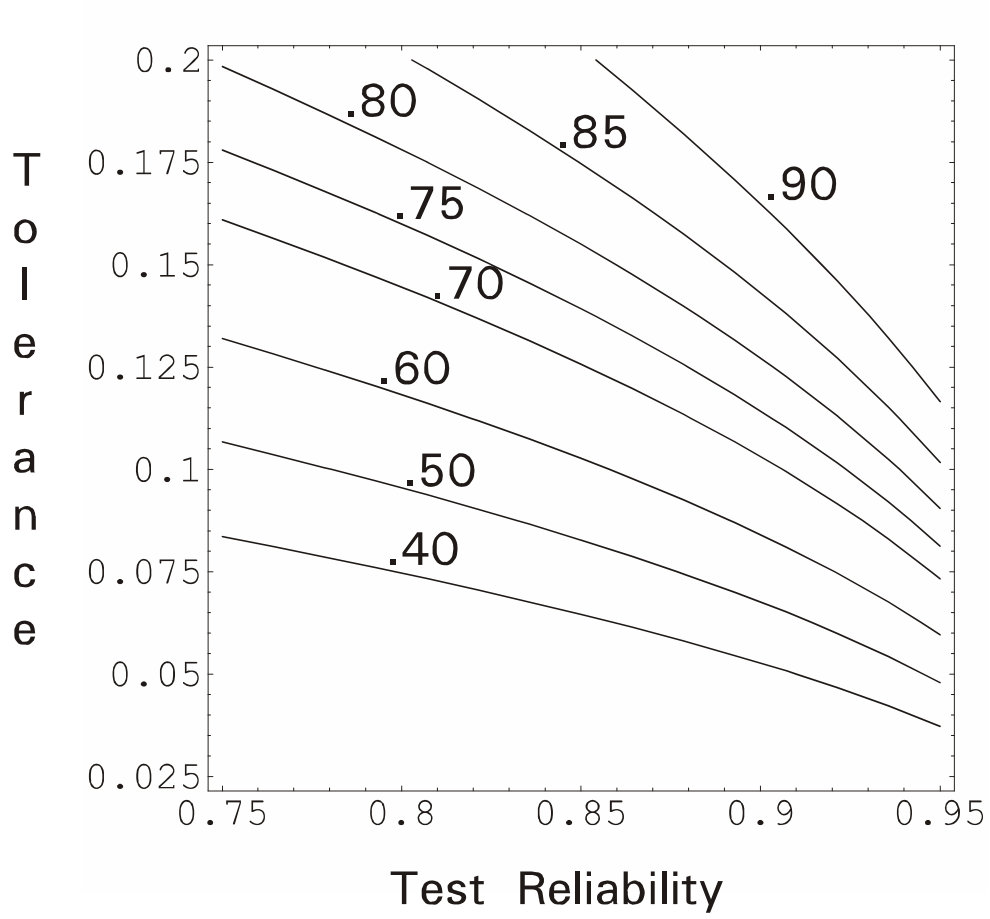


Exhibit IC
Plots of Hit-rate for students across various percentiles

Figure IC1. Hit-rate versus tolerance for $G_1(\tau)$ values .50, .60, .70, .80, .90; the three frames have reliability values .90 (a), .85 (b), .80 (c).

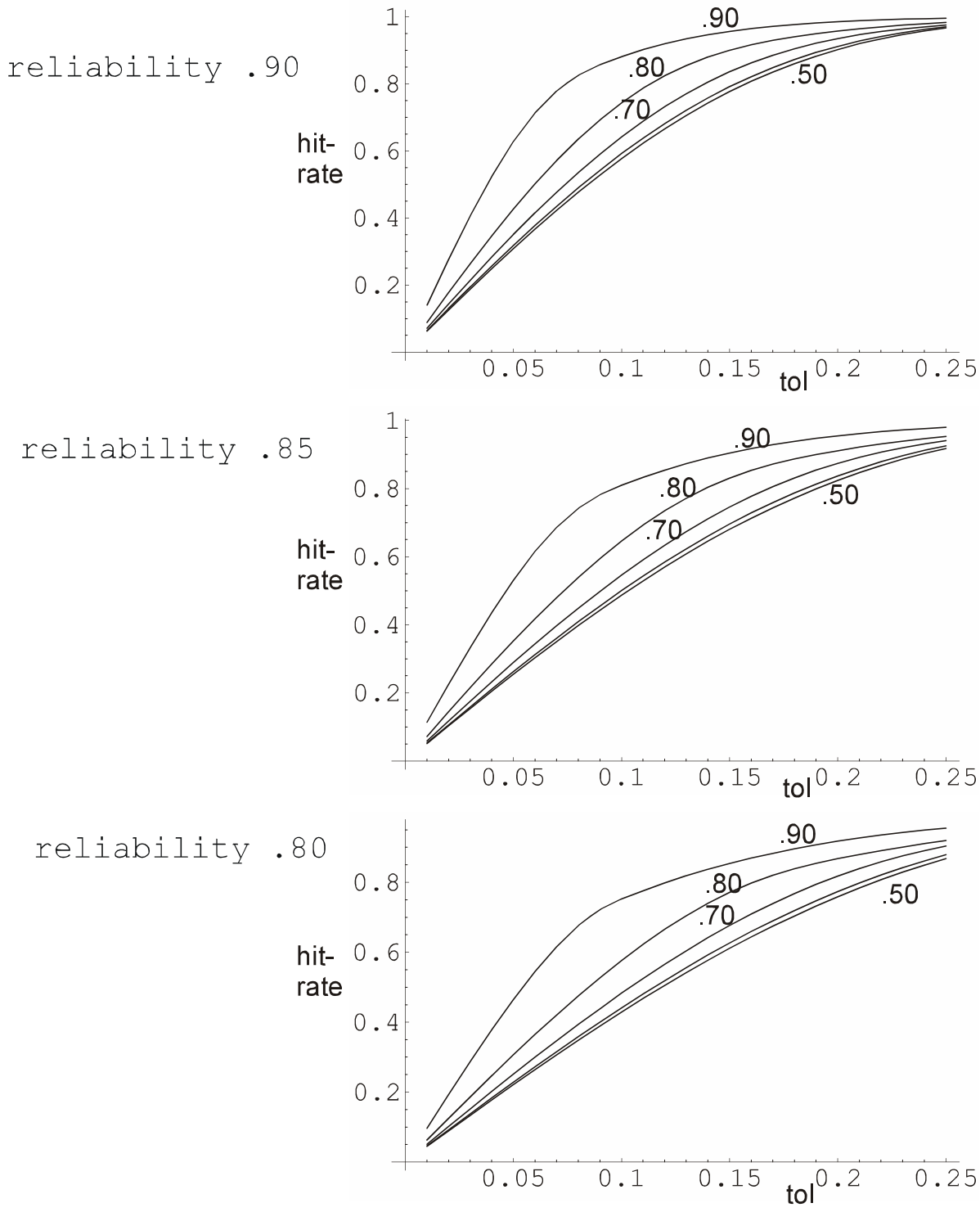


Exhibit IC continued

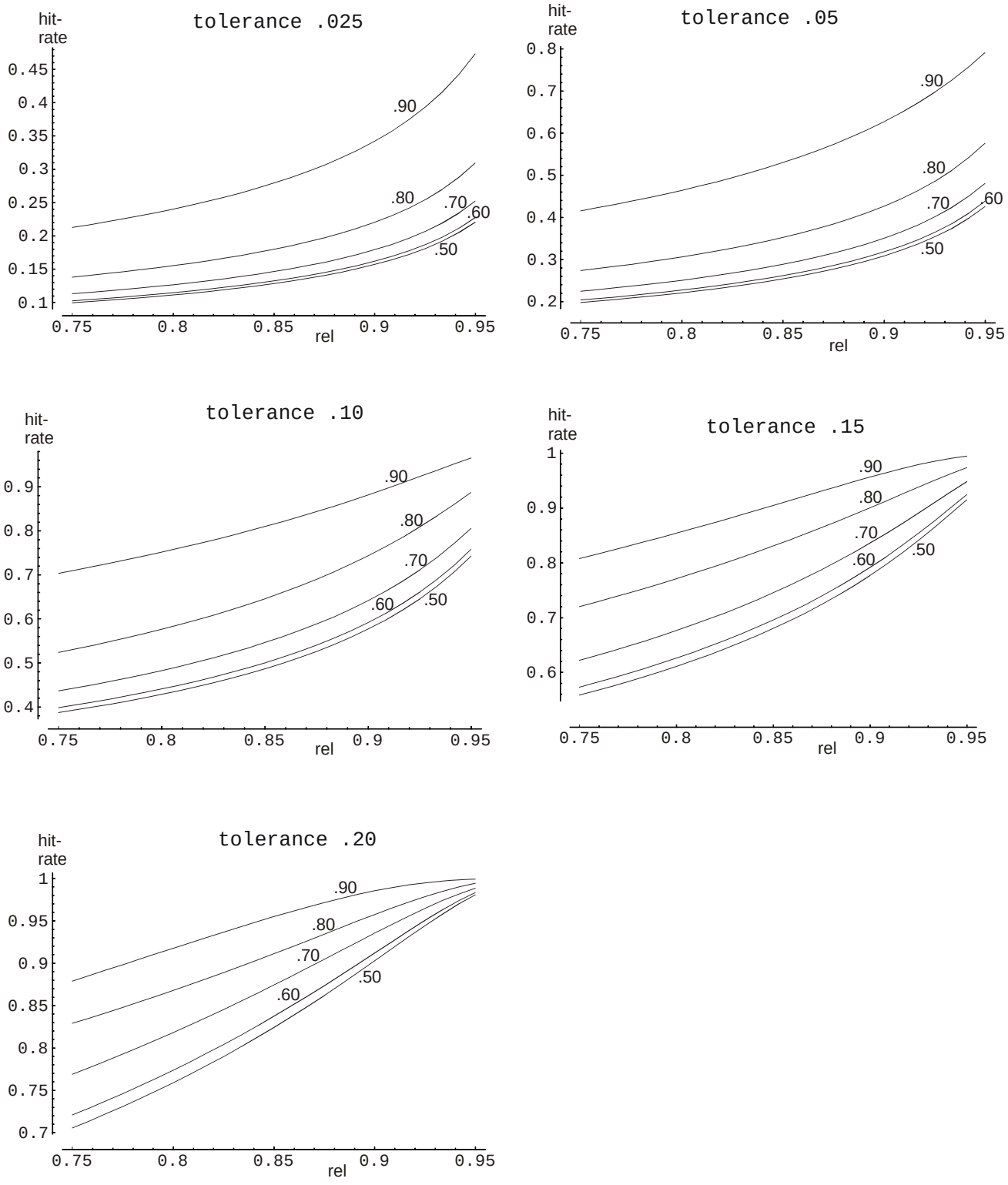


Figure IC2. Hit-rate versus reliability for labeled $G_1(\tau)$ values .50, .60, .70, .80, .90; each frame has a labeled value of tolerance.

Exhibit ID Hit-rate for $G_1(\tau)$ values .60, .70, .80, .90;

Table ID1. Hit-rate as a function of reliability (.7, .95), and tolerance (.025, .25) for $G_1(\tau)$ values .60, .70, .80, .90.

$G_1(t)$ = .60

	tolerance									
rel	.025	.05	.075	.10	.125	.15	.175	.20	.225	.25
0.70	0.0938	0.187	0.278	0.366	0.451	0.532	0.607	0.677	0.74	0.796
0.725	0.098	0.195	0.29	0.382	0.469	0.552	0.628	0.698	0.761	0.815
0.75	0.103	0.204	0.303	0.399	0.489	0.574	0.651	0.721	0.783	0.836
0.775	0.108	0.215	0.319	0.418	0.512	0.598	0.677	0.746	0.806	0.857
0.8	0.115	0.228	0.337	0.441	0.538	0.626	0.705	0.774	0.832	0.879
0.825	0.123	0.243	0.359	0.468	0.568	0.658	0.737	0.804	0.859	0.902
0.85	0.133	0.262	0.386	0.5	0.604	0.696	0.774	0.837	0.888	0.926
0.875	0.145	0.286	0.419	0.541	0.648	0.74	0.815	0.874	0.918	0.949
0.9	0.162	0.318	0.463	0.592	0.702	0.792	0.861	0.912	0.947	0.97
0.925	0.187	0.364	0.524	0.661	0.77	0.853	0.912	0.951	0.974	0.987
0.95	0.228	0.438	0.618	0.758	0.859	0.924	0.963	0.983	0.993	0.997

$G_1(t)$ = .70

	tolerance									
rel	.025	.05	.075	.10	.125	.15	.175	.20	.225	.25
0.70	0.103	0.205	0.305	0.401	0.493	0.579	0.657	0.727	0.786	0.835
0.725	0.108	0.215	0.318	0.418	0.512	0.599	0.678	0.747	0.805	0.851
0.75	0.113	0.225	0.333	0.437	0.534	0.622	0.701	0.769	0.825	0.868
0.775	0.12	0.237	0.351	0.458	0.558	0.648	0.727	0.793	0.846	0.886
0.8	0.127	0.251	0.371	0.483	0.585	0.676	0.754	0.818	0.868	0.904
0.825	0.136	0.268	0.395	0.512	0.617	0.709	0.785	0.845	0.891	0.922
0.85	0.147	0.289	0.424	0.546	0.654	0.746	0.819	0.874	0.914	0.941
0.875	0.161	0.316	0.46	0.588	0.698	0.788	0.856	0.905	0.938	0.959
0.9	0.179	0.351	0.507	0.642	0.752	0.836	0.896	0.936	0.96	0.975
0.925	0.207	0.401	0.572	0.711	0.817	0.891	0.937	0.965	0.98	0.989
0.95	0.252	0.481	0.668	0.806	0.896	0.948	0.975	0.989	0.995	0.998

$G_1(t)$ = .80

	tolerance							
rel	.025	.05	.075	.10	.125	.15	.175	.20
0.70	0.126	0.25	0.369	0.483	0.587	0.677	0.748	0.795
0.725	0.132	0.261	0.386	0.502	0.608	0.698	0.767	0.811
0.75	0.138	0.274	0.404	0.524	0.631	0.721	0.787	0.829
0.775	0.146	0.289	0.425	0.549	0.657	0.745	0.808	0.848
0.8	0.155	0.307	0.449	0.577	0.686	0.771	0.831	0.868
0.825	0.167	0.327	0.477	0.609	0.718	0.8	0.855	0.889
0.85	0.18	0.353	0.51	0.646	0.754	0.831	0.88	0.911
0.875	0.197	0.385	0.552	0.69	0.795	0.865	0.907	0.934
0.9	0.221	0.427	0.604	0.744	0.841	0.901	0.935	0.958
0.925	0.254	0.485	0.674	0.809	0.892	0.938	0.963	0.979
0.95	0.31	0.576	0.77	0.888	0.947	0.974	0.988	0.994

$G_1(t)$ = .90

	tolerance							
rel	.025	.05	.075	.10	.125	.15	.175	.20
0.70	0.192	0.378	0.547	0.663	0.719	0.766	0.807	0.841
0.725	0.202	0.396	0.569	0.683	0.739	0.787	0.827	0.86
0.75	0.213	0.416	0.592	0.704	0.761	0.808	0.847	0.879
0.775	0.226	0.439	0.619	0.727	0.784	0.831	0.868	0.898
0.8	0.24	0.465	0.648	0.752	0.809	0.854	0.89	0.918
0.825	0.258	0.495	0.68	0.78	0.836	0.879	0.912	0.937
0.85	0.279	0.53	0.716	0.81	0.864	0.905	0.934	0.955
0.875	0.307	0.573	0.757	0.844	0.895	0.931	0.955	0.972
0.9	0.342	0.627	0.804	0.882	0.927	0.957	0.975	0.986
0.925	0.393	0.697	0.858	0.923	0.959	0.979	0.99	0.995
0.95	0.473	0.791	0.919	0.966	0.986	0.995	0.998	0.999

Exhibit ID continued

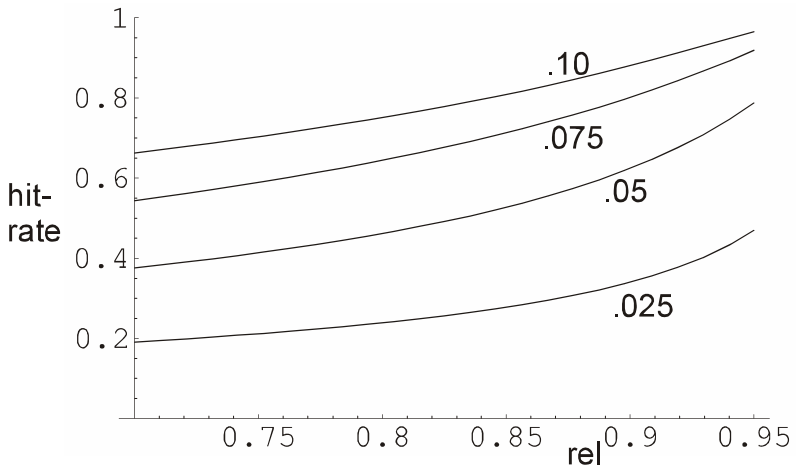


Figure ID1. Plot of hit-rate vs test reliability for student at 90th percentile; labeled tolerance values: {tol, .025, .10, .025}.

Figure ID2. Array of contour plot of hit-rate as a function of tolerance and test reliability for $G_1(\tau)$ values .60, .70, .80, .90; Hit-rate contours {.4,.5,.6,.7,.75,.8,.85, .9}.

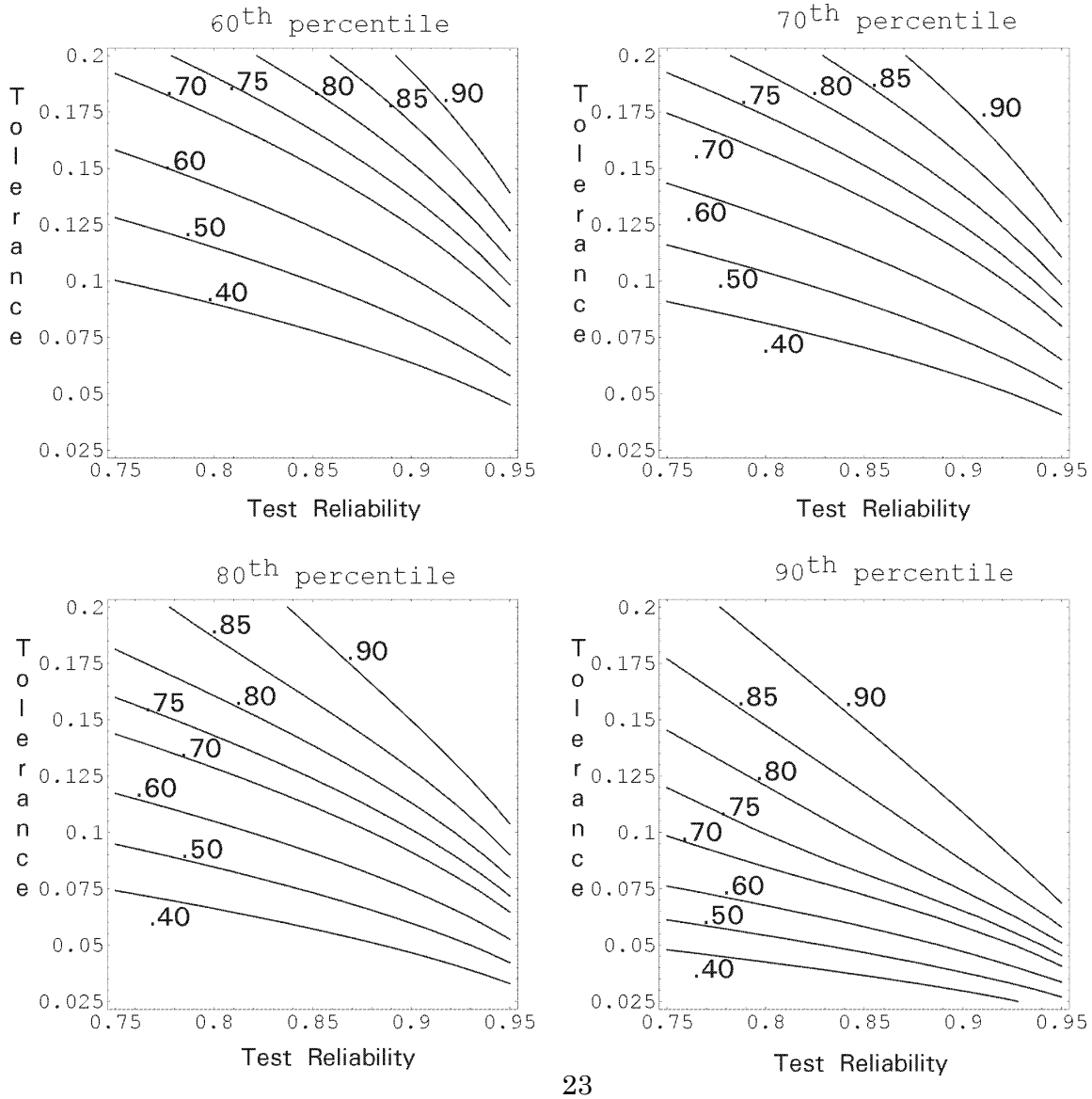


Exhibit IE
Percentile Accuracy averaged over all percentiles

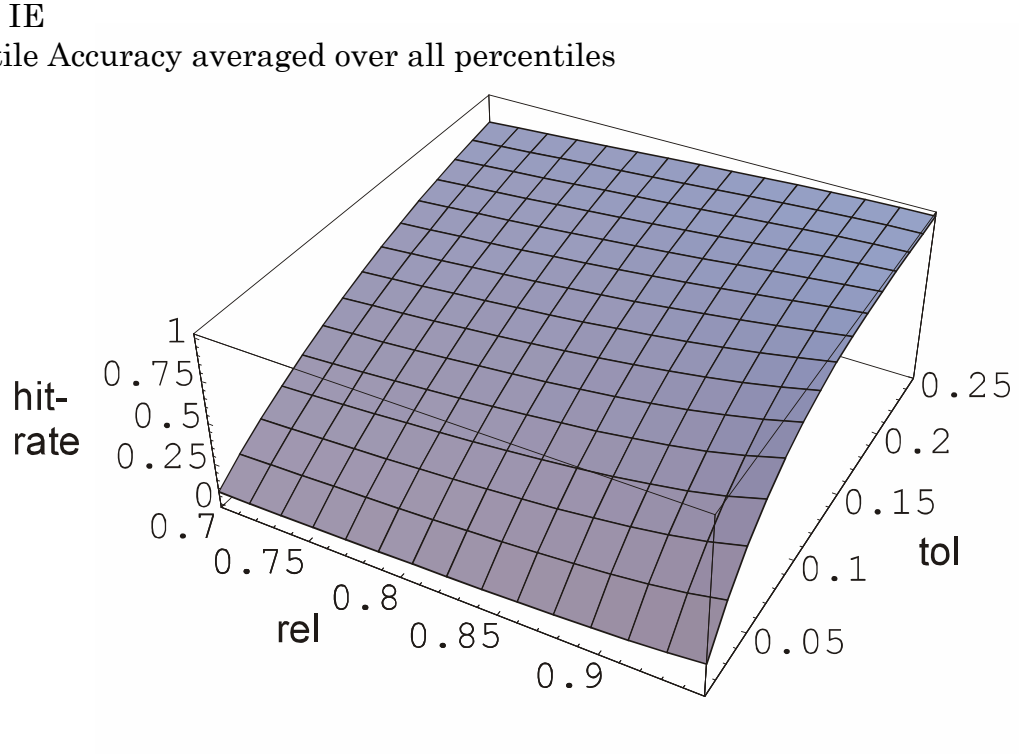


Figure IE1. 3D plot of hit-rate as a function of reliability (.7, .95), and tolerance (.025, .25) averaged over all percentiles.

Table IE1. Hit-rate as a function of reliability (.7, .95), and tolerance (.025, .25) averaged over all percentiles.

	tolerance							
	.025	.05	.075	.10	.125	.15	.175	.20
rel								
0.75	0.181	0.321	0.439	0.539	0.624	0.696	0.757	0.808
0.775	0.191	0.337	0.459	0.561	0.647	0.719	0.779	0.829
0.8	0.202	0.355	0.481	0.586	0.673	0.744	0.803	0.851
0.825	0.216	0.376	0.507	0.614	0.701	0.772	0.829	0.874
0.85	0.232	0.401	0.537	0.646	0.733	0.803	0.857	0.898
0.875	0.251	0.432	0.573	0.684	0.77	0.837	0.886	0.923
0.9	0.277	0.47	0.617	0.729	0.813	0.874	0.918	0.948
0.925	0.313	0.522	0.675	0.786	0.863	0.916	0.951	0.972
0.95	0.368	0.599	0.755	0.858	0.922	0.96	0.98	0.991

Exhibit IE continued

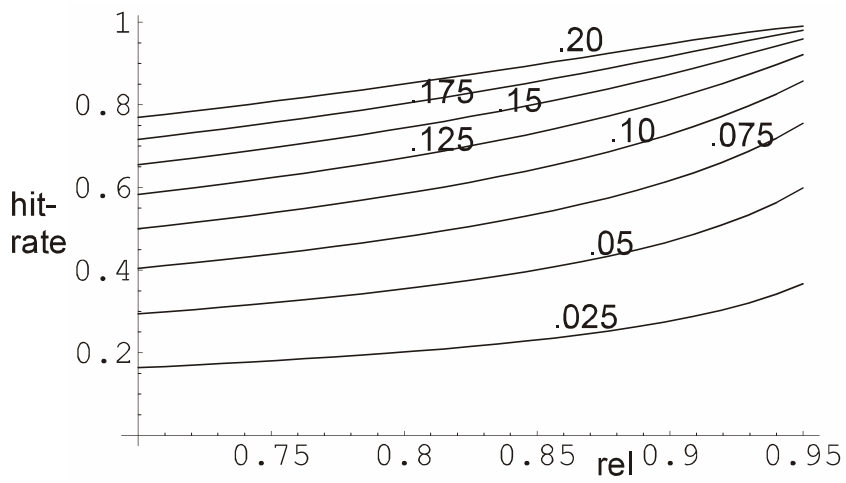
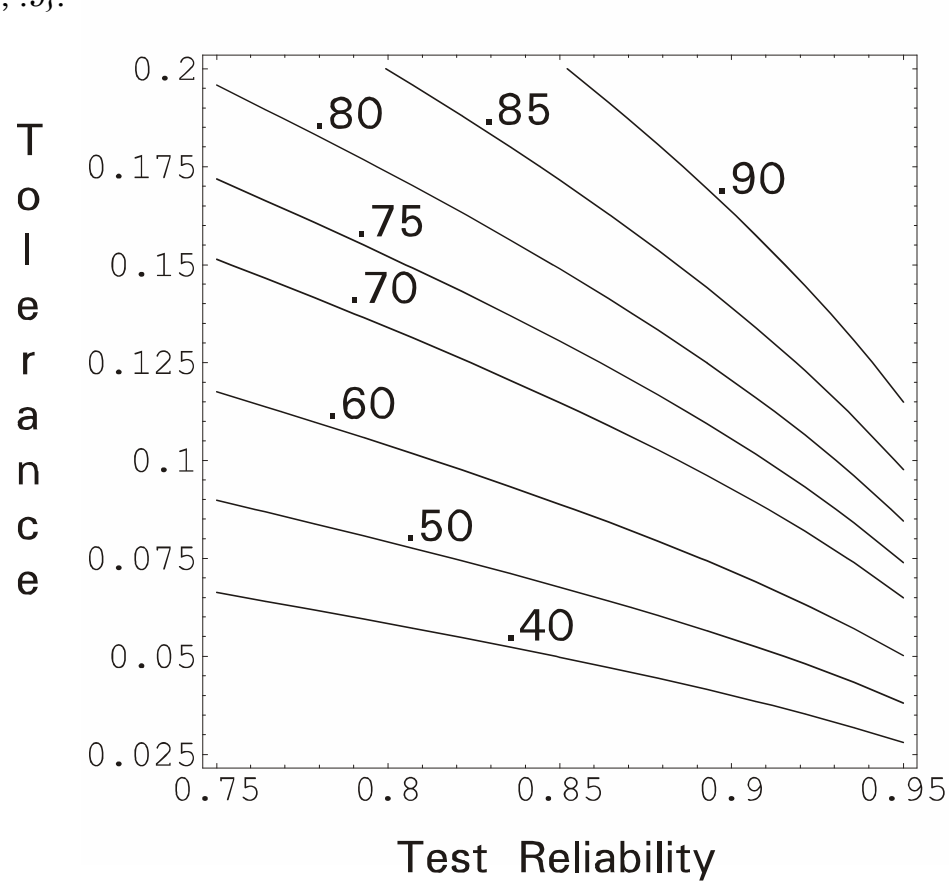


Figure IE2. Plot of hit-rate as a function of reliability (.7, .95), at each level of tolerance (.025, .05, .075, .10, .125, .15, .175, .20) averaged over all percentiles.

Figure IE3. Contour plot of hit-rate as a function of tolerance and test reliability averaged over all percentiles (randomly sampled student). Contours -> {.4, .5, .6, .7, .75, .8, .85, .9}.



A small aside: Percentile Bands, Score Reports. A common practice in standardized test reports is to display the uncertainty in the percentile rank score using ±1 s.e.m. confidence bands about the observed score S; this confidence band ranges from 100 G(S – s.e.m.) to 100 G(S + s.e.m.). For example, Harcourt Educational Measurement (HEM) reports "National Grade Percentile Bands" in the *Student Report* as shown in Figure 6.

SUBTESTS AND TOTALS	No. of Items	Raw Score	Scaled Score	National PR-S	National NCE
Total Reading	84	71	671	78-7	66.3
Vocabulary	30	27	678	78-7	66.3
Reading Comp.	54	44	668	75-6	64.2
Total Mathematics	78	68	676	90-8	77.0
Problem Solving	48	41	669	87-7	73.7
Procedures	30	27	686	87-7	73.7
Language	48	29	621	43-5	46.3
Lang Mechanics	24	10	590	20-3	32.3
Lang Expression	24	19	656	73-6	62.9
Spelling	30	20	629	54-5	52.1

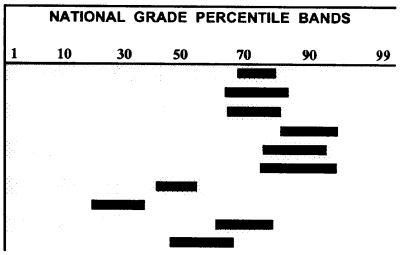


Figure 6. "National Grade Percentile Bands" in the SAT9 Student Report.

Assuming that this confidence band is intended as an interval for $G(\tau)$, a probability statement corresponding to the percentile band can be written as:

$$\Pr\{ G(S - \text{s.e.m.}) \le G(\tau) \le G(S + \text{s.e.m.}) \} = .683.$$

The probability statement for hit-rate in Equation 3.3 can be rewritten in the form

$$\Pr\{ G(S) - \text{tolerance} \le G(\tau) \le G(S) + \text{tolerance} \} = \text{hit-rate}.$$

So then selecting a hit-rate of .683 (i.e., a contour just below the .70 level) an approximate equivalence can be stated in terms of:

$$\begin{aligned} \text{tolerance} &\approx G(S + \text{s.e.m.}) - G(S) \text{ and/or} \\ &\approx G(S) - G(S - \text{s.e.m.}), \\ &\approx [G(S + \text{s.e.m.}) - G(S - \text{s.e.m.})]/2 . \end{aligned}$$

A different form of the approximate equivalence yields tol $\approx G(\tau) - G(\tau -$ s.e.m.) etc. (It's only for G(S) = .50 that both sides of that equivalence can be satisfied exactly, but that's not critical.) As the s.e.m. is proportional to $[(1 - \text{rel})/\text{rel}]^{1/2}$, for a stated test reliability the tolerance-equivalent needed to make the hit-rate = .683 in the probability statement can be computed as a function of test reliability. The entries in Table 1 are the approximate tolerance equivalence in the confidence interval probability statement (coverage .683) for test reliability (.7, .95) for scores at the 50^{th} , 75^{th} , and 90^{th} percentiles.

Table 1: Hit-rate and confidence interval equivalence for tolerance values.

	Score Percentile		
	50	75	90
rel			
0.70	0.208	0.180	0.124
0.75	0.191	0.164	0.109
0.8	0.173	0.146	0.0934
0.85	0.151	0.125	0.0776
0.9	0.124	0.102	0.0607
0.95	0.0885	0.0715	0.0411

To show the correspondence between the hit-rate probability statements used here and

the confidence bands in the score reports, find in Table IA1 a tolerance value close to the entry in the Table 1 .50 column with corresponding reliability and see if the resulting Table IA1 hit-rate with corresponding reliability is close to the confidence interval coverage .683. Note the following values for Table IA1 starting with the 50 score percentile column in Table 1: in Table IA1, rel .90, tol. .125, has hit-rate .686; rel .85, tol. .15, has hit-rate .68; rel .725, tol .20, has hit-rate .683. Using the 75 column of Table 1 and moving to Table IB1: rel .85, tol. .125, has hit-rate .697, and rel .90, tol .10, has hit-rate .685. Moreover, using the 90 percentile column of Table 1 and moving to Table ID1 for 90^{th} percentile rel .925, tol. .05, has hit-rate .697.

3.2 Comparing a Student Score to a Standard

A secondary set of calculations addresses the accuracy of an individual percentile rank score by examining the probability that a student's score is above a standard. There are many different treatments of this type of problem in the literature; calculations here focus on a standard set as a percentile of the observed norms distribution. Denote this percentile by P , and thus the standard is $G^{-1}(P)$. Given the student's percentile rank under perfect measurement, $G_1(\tau)$, calculate

$$\Pr\{G(S) > \text{standard}\} = \Pr\{S > G^{-1}(P)\} = 1 - \Pr\{S \leq G^{-1}(P)\} =$$

$$= 1 - \Phi[\{\Phi^{-1}[P] - (\sqrt{rel})\,\Phi^{-1}[G_1(\tau)]\}/(1 - rel)^{1/2}] \qquad (3.4)$$

If, instead, the student is identified by the percentile rank of τ in the observed norms, $G(\tau)$, the calculation yields:

$$\Pr\{S > G^{-1}(P)\} = 1 - \Phi[\{\Phi^{-1}[P] - \Phi^{-1}[G(\tau)]\}/(1 - rel)^{1/2}] \qquad (3.5)$$

Three sets of numerical illustrations for Equation 3.4 follow. The standard is chosen to be observed score norm 50^{th} percentile in Exhibit IF, the 90^{th} percentile in Exhibit IG, and the 75^{th} percentile in Exhibit IH. For each of these artificial standards used in the Exhibits, the quantities displayed are: (i) tabled values of Probability above Standard as a function of test reliability and true student percentile rank, $G_1(\tau)$, and (ii) a plot of Probability above Standard as a function of test reliability labeled with values of true student percentile rank, $G_1(\tau)$. For example, for test reliability .85, a student with $G_1(\tau) = .3$, has a 10.6% chance of obtaining a test score above the 50^{th} percentile, or for test reliability .90, a student with $G_1(\tau) = .6$, has a 22.4% chance of obtaining a test score below the 50^{th} percentile, (Exhibit IF). Furthermore, for test reliability .85, a student with $G_1(\tau) = .90$ (e.g., a potential GATE student), has a 9.95% chance of obtaining a test score below the 75^{th} percentile (Exhibit IH).

Insert Exhibits IF, IG, IH here

Exhibit IF
Comparing a Student to a Standard:
Standard Set at Observed Norms 50th percentile

Table IF1. Probability above Standard for student with true percentile rank as a function of test reliability.

	$G_1(\tau)$						
	.2	.25	.3	.35	.4	.45	.5
rel							
0.75	0.0725	0.121	0.182	0.252	0.33	0.414	0.5
0.775	0.0591	0.105	0.165	0.237	0.319	0.408	0.5
0.8	0.0462	0.0887	0.147	0.22	0.306	0.401	0.5
0.825	0.0338	0.0715	0.127	0.201	0.291	0.392	0.5
0.85	0.0226	0.0542	0.106	0.18	0.273	0.382	0.5
0.875	0.013	0.0372	0.0827	0.154	0.251	0.37	0.5
0.9	0.00579	0.0215	0.0578	0.124	0.224	0.353	0.5
0.925	0.00156	0.00892	0.0328	0.088	0.187	0.329	0.5
0.95	0.000122	0.00164	0.0111	0.0465	0.135	0.292	0.5

	$G_1(\tau)$					
	.55	.6	.65	.7	.75	.8
rel						
0.75	0.586	0.67	0.748	0.818	0.879	0.928
0.775	0.592	0.681	0.763	0.835	0.895	0.941
0.8	0.599	0.694	0.78	0.853	0.911	0.954
0.825	0.608	0.709	0.799	0.873	0.928	0.966
0.85	0.618	0.727	0.82	0.894	0.946	0.977
0.875	0.63	0.749	0.846	0.917	0.963	0.987
0.9	0.647	0.776	0.876	0.942	0.978	0.994
0.925	0.671	0.813	0.912	0.967	0.991	0.998
0.95	0.708	0.865	0.953	0.989	0.998	1.

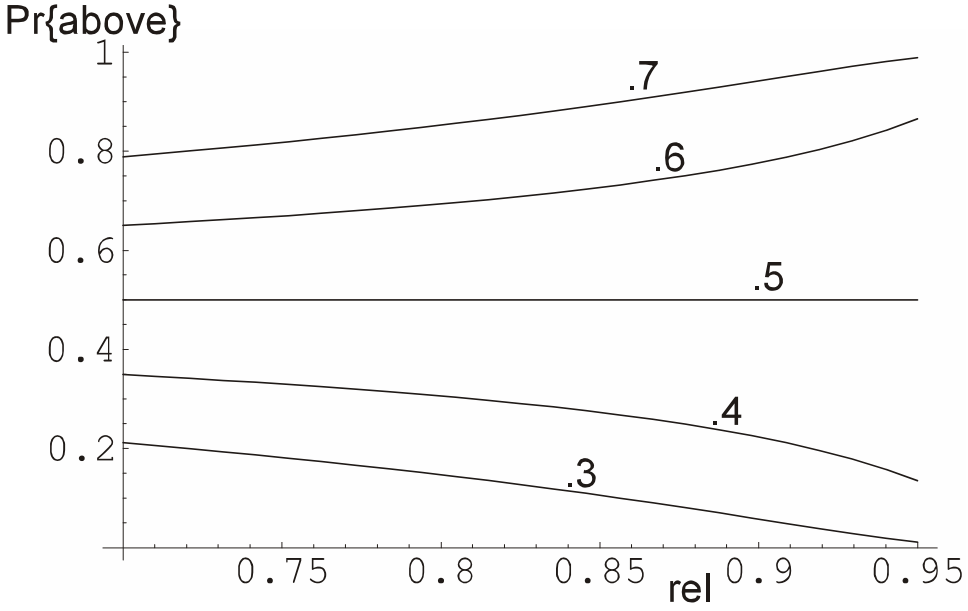


Figure IF1. Probability above Standard as a function of test reliability; each curve is labeled with $G_1(\tau)$ value (student true percentile rank).

Exhibit IG
Comparing a Student to a Standard:
Standard Set at Observed Norms 90th percentile

Table IG1. Probability above Standard for student with true percentile rank as a function of test reliability.

	$G_1(\tau)$					
	.65	.675	.70	.725	.75	.775
rel						
0.75	0.029	0.0378	0.049	0.0633	0.0815	0.105
0.775	0.0235	0.0315	0.0419	0.0557	0.0735	0.0968
0.8	0.0181	0.0251	0.0346	0.0474	0.0647	0.0877
0.825	0.013	0.0188	0.0271	0.0387	0.0549	0.0773
0.85	0.00839	0.0129	0.0197	0.0296	0.0443	0.0654
0.875	0.00459	0.00767	0.0126	0.0205	0.0329	0.052
0.9	0.00189	0.00356	0.00658	0.0119	0.0212	0.037
0.925	0	0.00101	0.00227	0.00494	0.0104	0.0213
0.95	0	0	0	0	0.00263	0.00737

	$G_1(\tau)$							
	.80	.825	.85	.875	.90	.925	.95	.975
rel								
0.75	0.134	0.172	0.221	0.284	0.366	0.472	0.613	0.797
0.775	0.127	0.167	0.218	0.285	0.373	0.488	0.637	0.825
0.8	0.119	0.16	0.214	0.286	0.381	0.505	0.664	0.854
0.825	0.108	0.151	0.208	0.286	0.389	0.525	0.694	0.883
0.85	0.0959	0.139	0.2	0.284	0.398	0.547	0.728	0.913
0.875	0.081	0.125	0.189	0.281	0.407	0.573	0.766	0.941
0.9	0.0633	0.106	0.173	0.274	0.418	0.605	0.811	0.966
0.925	0.0424	0.0811	0.149	0.261	0.429	0.647	0.864	0.986
0.95	0.0196	0.0487	0.112	0.237	0.442	0.707	0.925	0.998

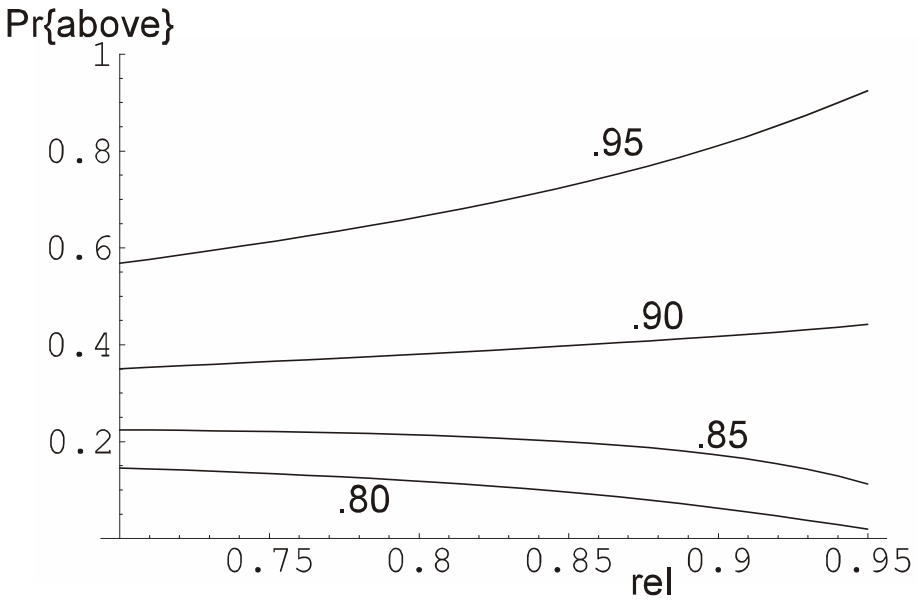


Figure IG1. Probability above Standard as a function of test reliability; each curve is labeled with $G_1(\tau)$ value (student true percentile rank).

Exhibit IH
Comparing a Student to a Standard:
Standard Set at Observed Norms 75th percentile

Table IH1. Probability above Standard for student with true percentile rank as a function of test reliability.

	$G_1(\tau)$							
	.50	.6	.65	.7	.75	.8	.85	.90
rel								
0.75	0.0887	0.181	0.248	0.33	0.428	0.543	0.672	0.808
0.775	0.0775	0.171	0.24	0.327	0.432	0.556	0.692	0.831
0.8	0.0658	0.158	0.23	0.323	0.437	0.569	0.714	0.854
0.825	0.0534	0.144	0.219	0.318	0.441	0.585	0.738	0.879
0.85	0.0408	0.127	0.205	0.311	0.446	0.603	0.766	0.905
0.875	0.0282	0.108	0.187	0.301	0.451	0.625	0.798	0.931
0.9	0.0165	0.0849	0.164	0.288	0.456	0.652	0.836	0.957
0.925	0.00689	0.0578	0.134	0.267	0.462	0.689	0.88	0.979
0.95	0.00128	0.0279	0.0906	0.233	0.47	0.743	0.933	0.995

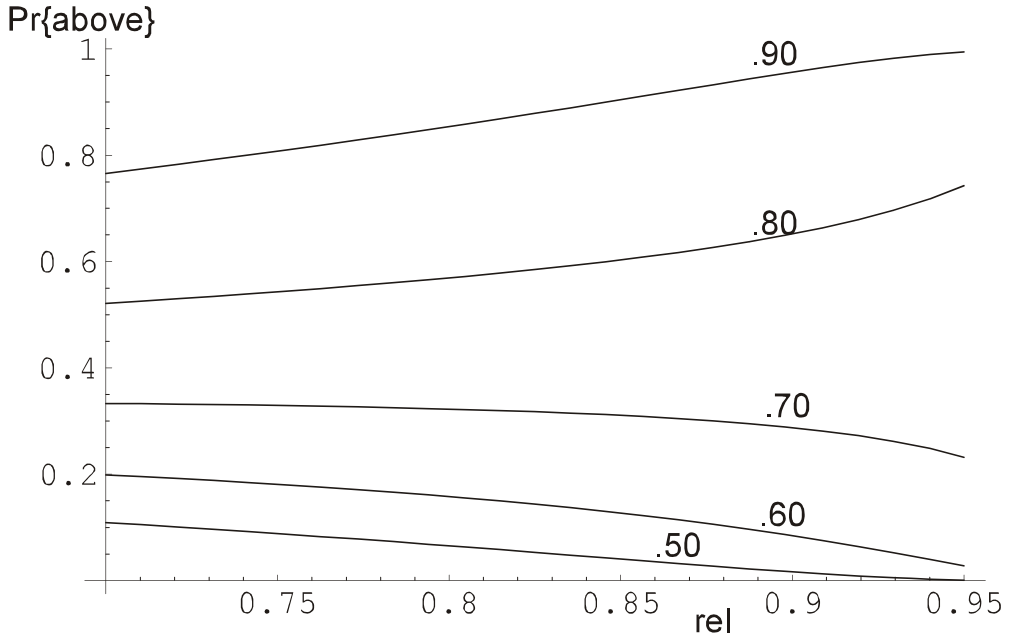


Figure IH1. Probability above Standard as a function of test reliability; each curve is labeled with $G_1(\tau)$ value (student true percentile rank).

4. Test-Retest Consistency for a Single Student

Another strategy for describing accuracy, test-retest consistency, uses only (potentially) observed percentile-rank scores. Consider two (contemporaneous) scores from a single student, denoted here by $G(S_a)$ and $G(S_b)$, and examine the size of the discrepancy between the scores, $|G(S_a) - G(S_b)|$. In terms of the depiction in Figure 3 the scores S_a and S_b are two draws from the labeled S-distribution. Test-retest comparisons have a long history in measurement, and the test-retest correlation is one of the standard approaches to estimating reliability. The calculations here, which describe test-retest consistency as a function of test reliability, allow us to ask, How well does a test-retest correlation inform about accuracy?

Some additional motivational references and scenarios. The Parent Assistance Packet from California Department of Education (CDE, 1999) gives the following caption for interpreting the National Percentile Rank Scores:

> "No single number can exactly represent a student's level of achievement. If a student were to take a different form of the test within a short period of time, that score could vary from the first score." (page TM-15).

The test-retest accuracy calculations here answer the question (under classical test theory assumptions), How close would two (contemporaneous) percentile rank scores be?

Another approach to test-retest consistency is to follow the amateur handyman dictum: "*measure twice, cut once*" (the title of Norm Abram's fine text, Abram, 1996). Represent accuracy as how close together (or far apart) two measurements on the same student would be; if you measured a board twice and the two measurements were not close, you may not be satisfied with the quality of your measurement. Another story for this same calculation is "identical twins separated at test-time." For example, two kids (e.g., next-door-neighbors) with identical achievement (both really belong at the same percentile). What's the chances of their percentile rank scores being more than 10 percentile points different?

Yet another motivation for these calculations is the recent history in California, in the Pupil Testing Incentive Program, of seeking comparability of individual scores across tests. To wit, AB265 in 1995 sought the reporting of "valid, reliable, comparable individual pupil results" from many different test publishers. A canonical example in the discussion of AB265 is that a student is in district A in fifth grade and moves to district B in sixth grade, and even though districts A and B have selected tests from two different publishers, successful comparability of the tests will allow the fifth-grade score and the sixth-grade score to be interpretable in terms of student progress. The calculations herein on test-retest consistency relate to the empirical consequences of having perfect linking between tests in the

following sense: The best results that could be seen from comparability would be the test-retest consistency seen from a student taking the same test. To take this into the realm of a thought experiment, a newspaper might choose to do the following study. Assemble a group of 100 "typical" students and pay them to spend one Saturday taking Publisher A test and the next Saturday Publisher B test (A and B being prime in the list of tests that are claimed to have been made comparable). The journalists may reasonably expect the successive scores for the same student to be pretty much the same.

For the calculations of test-retest consistency select a $G_1(\tau)$, and for that $G_1(\tau)$ compute:

$$\begin{aligned} \text{retest} &= \Pr\{\,|\,G(S_a) - G(S_b)\,|\,\le \text{ tolerance}\,|\,G_1(\tau)\,\} \\ &= \Pr\{G(S_a) - G(S_b) \le \text{tol}\,\} - \Pr\{G(S_a) - G(S_b) < -\text{tol}\,\} \end{aligned} \tag{4.1}$$

For example, for a student with $G_1(\tau) = .50$ (true standing at 50th percentile), obtain two test scores (S_a and S_b) and calculate $\Pr\{|\,G(S_a) - G(S_b)\,| < .10\}$. Table IIA1 in Exhibit IIA shows that quantity to be .492 for test reliability .925 (i.e., even for this high reliability slightly more than half of the test-retest pairs will be more than 10 percentile points apart). For test reliability .85, only 36.7% of the test-retest pairs will be within 10 percentile points (and less than two-thirds will be within 20 percentile points).

Computation of Retest Probability: Technical Details

The computation of the retest probability is implemented using the following conditioning argument. For a student with a specified $G_1(\tau)$, condition on a draw of an S_b from the S-distribution ($S\,|\,\tau \sim N[\,\tau\,,\,\sigma_N(1-\text{rel})^{1/2}\,]$) and express that S_b in terms of its fractile of the S-distribution, ps_b, to obtain:

$$\Pr\{G(S_a) - G(S_b) \le \text{tolerance} \mid ps_b\} = \Pr\{S_a \le G^{-1}[G(S_b) + \text{tol}] \mid ps_b\} =$$

$$\Phi\Big[\{\Phi^{-1}[\Phi[(1-\text{rel})^{1/2}\,\Phi^{-1}[ps_b] + (\sqrt{\text{rel}})\,\Phi^{-1}[G_1(\tau)]\,] + \text{tol}\,] - (\sqrt{\text{rel}})\,\Phi^{-1}[G_1(\tau)]\,\}/(1-\text{rel})^{1/2}\Big]$$

Then uncondition by integrating over ps_b in [0,1] the quantity $\Pr\{G(S_a) - G(S_b) \le \text{tolerance} \mid ps_b\} - \Pr\{G(S_a) - G(S_b) < -\text{tolerance} \mid ps_b\}$:

$$\begin{aligned} \text{retest} = \int_0^1 \Big[&(\Phi[\{\Phi^{-1}[\Phi[(1-\text{rel})^{1/2}\,\Phi^{-1}[ps_b] + (\sqrt{\text{rel}})\,\Phi^{-1}[G_1(\tau)]\,] + \text{tol}\,] - \\ &(\sqrt{\text{rel}})\,\Phi^{-1}[G_1(\tau)]\,\}/(1-\text{rel})^{1/2}]) - \\ &(\Phi[\{\Phi^{-1}[\Phi[(1-\text{rel})^{1/2}\,\Phi^{-1}[ps_b] + (\sqrt{\text{rel}})\,\Phi^{-1}[G_1(\tau)]\,] - \text{tol}\,] - \\ &(\sqrt{\text{rel}})\,\Phi^{-1}[G_1(\tau)]\,\}/(1-\text{rel})^{1/2}])\Big]\, d\,ps_b \end{aligned} \tag{4.2}$$

Results for 50th , 25th , 75th percentiles. Exhibits IIA ($G_1(\tau) = .50$) and IIB ($G_1(\tau) = .75, .25$) have the same structure: a Table whose entries give numerical values for dependence of the retest probability on test reliability and the value of tolerance. The dependence of retest probability on test reliability is most directly shown in the Figure below the table; each plotted curve is labeled with a specified level for the tolerance. The final entry in the exhibit is the contour plot; each of the contours is labeled with a specified level for the retest probability. The purpose of the contour plot is to show the combination of test reliability and tolerance needed to achieve a specific level for the retest probability. As noted for Section 3, the 50th percentile ($G_1(\tau) = .50$) produces the worst results for hit-rate (even though the measurement error variance is the same for all score values, because the percentile rank scores depend on G(S) which is bounded above and flattens out for higher and lower percentiles). Taking the ratio of the entries of Table IIB1 to those of Table IIA1, the retest probability for $G_1(\tau) = .75$ is about 1.2 times as large as that for $G_1(\tau) = .50$ for the lower values of tolerance, the ratio decreasing to about 1.05 for high reliability and larger tolerance values (quite similar to the corresponding comparison noted in Section 3). Also of note are comparisons between the retest probability and percentile discrepancy hit-rate in Section 3; roughly, the retest probability is about 3/4 to 4/5 as large for comparing Exhibits IIA and IA. For $G_1(\tau) = .50$, the ratio of the entries of Table IIA1 to those of Table IA1 is near .75 for the lower values of tolerance, the ratio increasing to the range .85 to .90 for high reliability and larger tolerance values. For $G_1(\tau) = .75$, the ratio of the entries of Table IIB1 to those of Table IB1 is similar for low tolerance value, but the ratio is larger (>.90) for high reliability and larger tolerance values.

Insert Exhibits IIA, IIB here

Other percentiles. The results in Exhibits IIA and IIB can be amplified by additional displays for other choices of $G_1(\tau)$; i.e., How much do the results in Exhibits IIA, IIB change for other choices of $G_1(\tau)$? Using test reliability .90, Exhibit IIC gives plots of the retest probability with each curve labeled for percentiles 50, 60, 70, 80, 90. (Only curves for $G_1(\tau) > .50$ are used for comparison as results for $G_1(\tau) = .2$ are identical to $G_1(\tau) = .8$ and so forth.) Plots of retest probability vs tolerance show 50th and 60th percentiles are almost indistinguishable; $G_1(\tau) = .80$ gives notably higher retest probabilities than $G_1(\tau) = .60$. The subsequent tables give the retest probability for test reliability and tolerance, for $G_1(\tau) = .60, .90$.

Insert Exhibit IIC here

Exhibit IIA
Test-Retest Accuracy for student at 50th percentile

Table IIA1. Retest probability as a function of reliability (.7, .95), and tolerance (.025, .25) for student at 50th percentile.

	tolerance					
	.05	.075	.10	.15	.20	.25
rel						
0.70	0.139	0.208	0.275	0.403	0.52	0.624
0.725	0.144	0.215	0.285	0.416	0.536	0.642
0.75	0.15	0.224	0.296	0.432	0.555	0.662
0.775	0.157	0.234	0.309	0.45	0.576	0.685
0.8	0.166	0.246	0.324	0.471	0.6	0.71
0.825	0.176	0.261	0.343	0.496	0.629	0.739
0.85	0.188	0.279	0.367	0.527	0.663	0.772
0.875	0.204	0.303	0.396	0.565	0.704	0.811
0.9	0.226	0.334	0.436	0.614	0.754	0.856
0.925	0.259	0.38	0.492	0.68	0.817	0.906
0.95	0.312	0.453	0.579	0.774	0.895	0.959

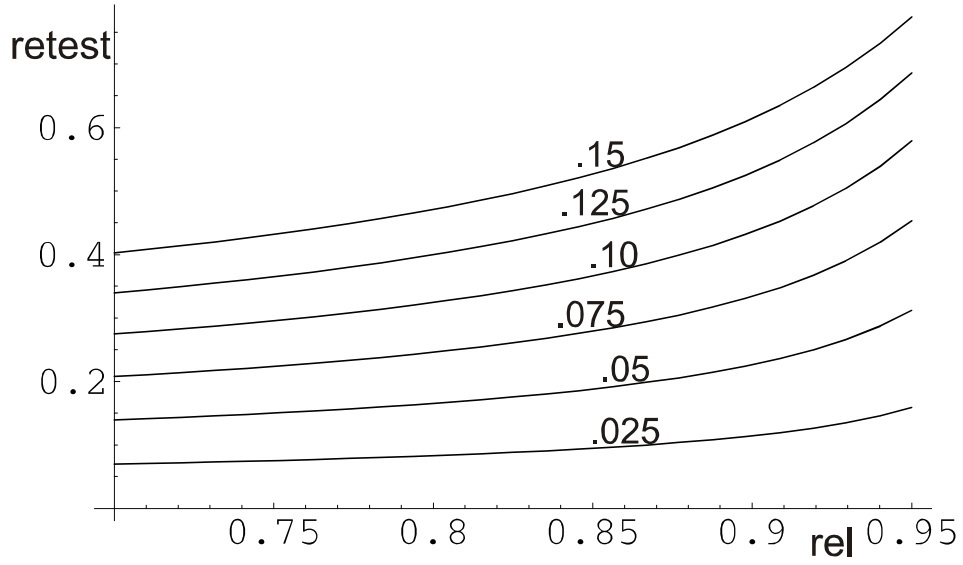


Figure IIA1. Plot of retest probability as a function of reliability (.7, .95), at each level of tolerance (.025, .05, .075, .10, .125, .15) for student at 50th percentile.

Exhibit IIA continued

Figure IIA2. Contour plot of retest probability as a function of tolerance and test reliability: Student at 50th percentile.

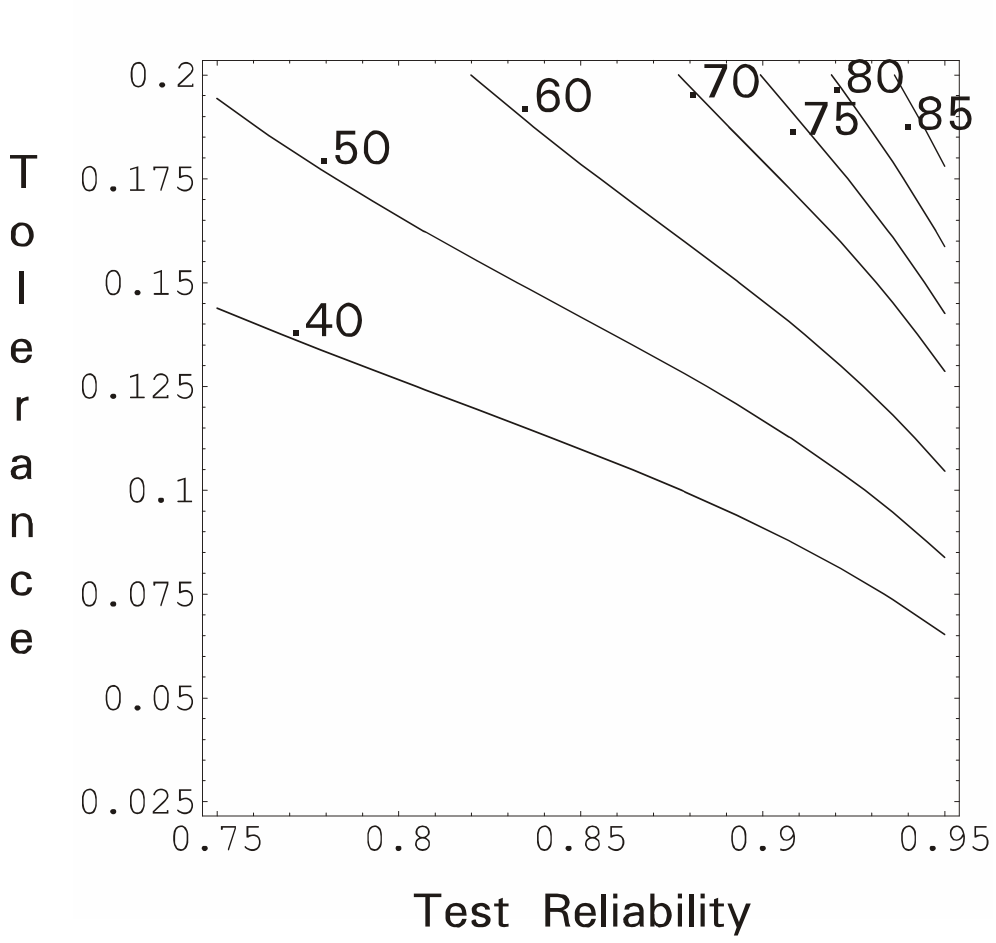


Exhibit IIB
Test-Retest Accuracy for student at 75th percentile

Table IIB1. Retest probability as a function of reliability (.7, .95), and tolerance (.025, .25) for student at 75th percentile.

	tolerance					
	.05	.075	.10	.15	.20	.25
rel						
0.70	0.167	0.248	0.325	0.467	0.59	0.694
0.725	0.174	0.257	0.337	0.484	0.61	0.714
0.75	0.182	0.269	0.352	0.503	0.632	0.736
0.775	0.191	0.282	0.369	0.525	0.656	0.76
0.8	0.201	0.297	0.388	0.551	0.684	0.787
0.825	0.214	0.316	0.412	0.58	0.715	0.816
0.85	0.23	0.339	0.44	0.615	0.751	0.848
0.875	0.251	0.368	0.476	0.658	0.792	0.883
0.9	0.278	0.406	0.522	0.711	0.84	0.92
0.925	0.318	0.461	0.586	0.778	0.895	0.956
0.95	0.383	0.547	0.682	0.864	0.952	0.986

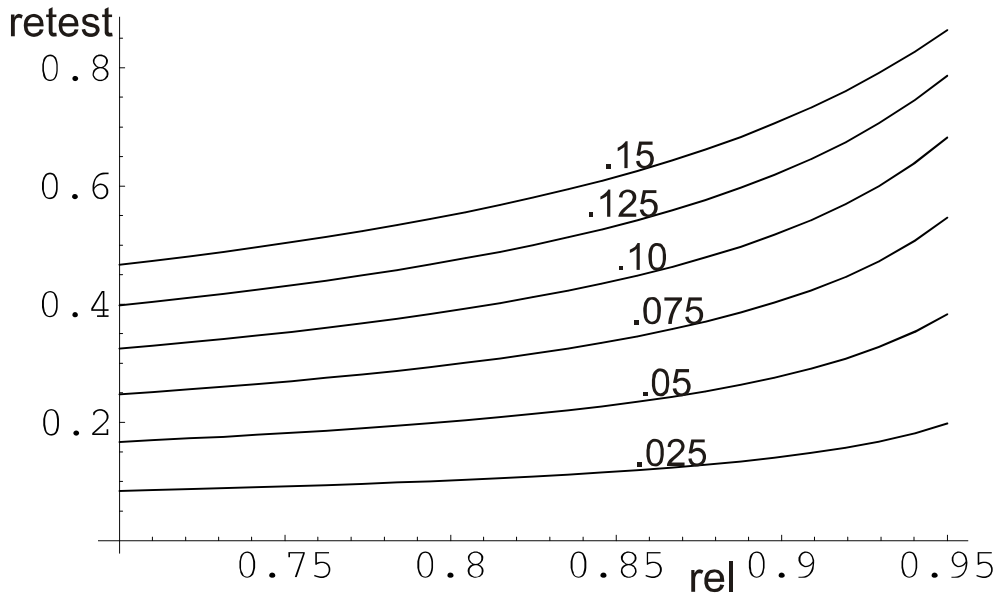


Figure IIB1. Plot of retest probability as a function of reliability (.7, .95), at each level of tolerance (.025, .05, .075, .10, .125, .15) for student at 75th percentile.

Exhibit IIB continued

Figure IIB2. Contour plot of retest probability as a function of tolerance and test reliability: Student at 75th percentile.

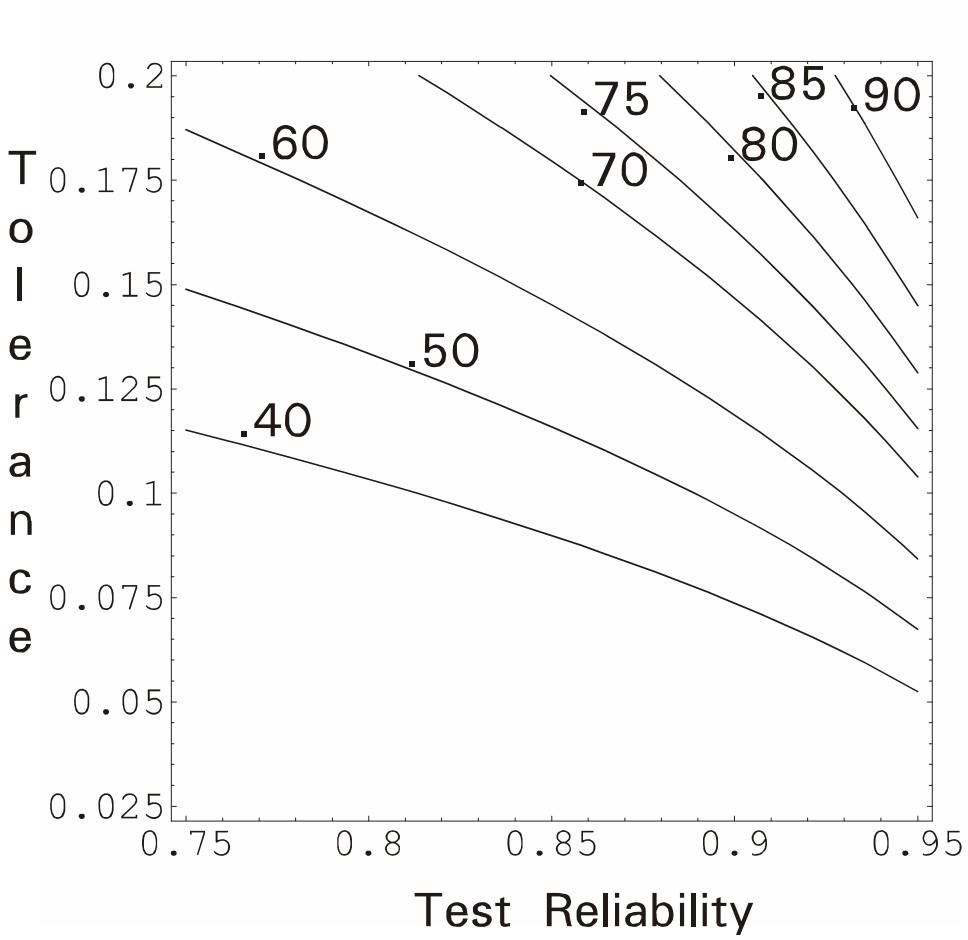


Exhibit IIC
Test-Retest Accuracy for additional percentiles

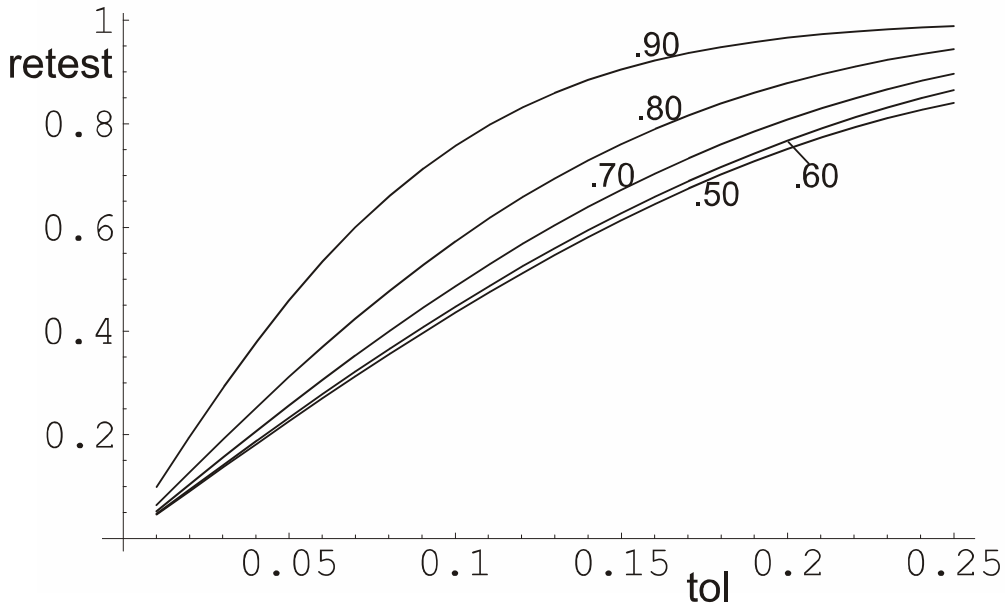


Figure IIC1. Plot of retest probability for reliability .90 as a function of tolerance (.025, .25) for student at labeled $G_1(\tau)$ values (.5, .6, .7, .8, .9).

Exhibit IIC continued

Table IIC1. Retest probability as a function of reliability (.7, .95), and tolerance (.025, .25) for student at 60th percentile.

	tolerance					
	.05	.075	.10	.15	.20	.25
rel						
0.70	0.143	0.213	0.281	0.411	0.53	0.635
0.725	0.148	0.221	0.292	0.426	0.547	0.653
0.75	0.154	0.23	0.303	0.442	0.566	0.673
0.775	0.162	0.24	0.317	0.46	0.587	0.696
0.8	0.17	0.253	0.333	0.482	0.612	0.722
0.825	0.181	0.268	0.352	0.508	0.642	0.751
0.85	0.194	0.287	0.376	0.539	0.676	0.784
0.875	0.21	0.311	0.407	0.578	0.717	0.822
0.9	0.233	0.344	0.447	0.628	0.768	0.866
0.925	0.266	0.391	0.505	0.695	0.83	0.915
0.95	0.321	0.466	0.593	0.788	0.905	0.964

Table IIC2. Retest probability as a function of reliability (.7, .95), and tolerance (.025, .25) for student at 90th percentile.

	tolerance					
	.05	.075	.10	.15	.20	.25
rel						
0.70	0.261	0.374	0.473	0.634	0.751	0.835
0.725	0.274	0.392	0.495	0.659	0.776	0.856
0.75	0.289	0.413	0.519	0.686	0.801	0.878
0.775	0.307	0.436	0.547	0.716	0.828	0.899
0.8	0.327	0.463	0.578	0.748	0.856	0.921
0.825	0.35	0.494	0.613	0.783	0.884	0.941
0.85	0.378	0.531	0.653	0.821	0.913	0.96
0.875	0.414	0.575	0.701	0.862	0.941	0.977
0.9	0.46	0.631	0.758	0.905	0.967	0.989
0.925	0.523	0.704	0.826	0.948	0.987	0.997
0.95	0.619	0.802	0.907	0.984	0.998	1.

5. Comparing Two Students

Another extension of the basic accuracy calculations is to consider two students who under perfect measurement have different score levels. For example, as is shown in Figure 4, take Student 1 at 50th percentile, and Student 2 at 75th percentile ($G_1(\tau_1) = .50$, $G_1(\tau_2) = .75$). From Figure 4, the depicted score distributions for S_1 and S_2 (for test reliability .90) overlap considerably, and each does span a wide range of values of G(Y). In Section 5.1 (and Exhibit IIIA), the focus is on calculating the probability of a reversal (e.g., for Student 1 and Student 2 $G(S_1) > G(S_2)$ even though $G_1(\tau_1) < G_1(\tau_2)$). In Section 5.2 (and Exhibit IIIB), the distribution of the difference of the two students' observed percentile rank scores is shown–e.g., $\Pr\{G(S_1) - G(S_2) \leq \text{bound}\}$.

5.1. Probability of Reversal

Results for probability of reversal are arrayed in Exhibit IIIA. The quantity labeled as reversal is:

$$\text{reversal} = \Pr\{ G(S_1) - G(S_2) > 0 \mid G_1(\tau_1), G_1(\tau_2)\}, \quad \text{where } G_1(\tau_1) < G_1(\tau_2) .$$

$$= \Phi[\{\Phi^{-1}[G_1(\tau_1)] - \Phi^{-1}[G_1(\tau_2)]\} \{\text{rel} /(2 \cdot (1 - \text{rel}))\}^{1/2}] \qquad (5.1)$$

The severity of a reversal depends on the magnitude of $G_1(\tau_1) - G_1(\tau_2)$; obviously if $G_1(\tau_1) = .50$ and $G_1(\tau_2) = .55$, some reasonable probability of reversal in the observed scores would be expected, even for accurate tests.

Five comparison scenarios are shown in Exhibit IIIA; the first two are $G_1(\tau_1) = .50$, $G_1(\tau_2) = .75$ and $G_1(\tau_1) = .75$, $G_1(\tau_2) = .90$. Tables and plots of reversal probability as a function of test reliability are shown for each scenario. In Scenario 3, $G_1(\tau_2) = G_1(\tau_1) + .20$, and reversal is tabled as a function of test reliability over a range of of $G_1(\tau_1)$ values, with a corresponding plot for test reliability .90. Additional tables are presented for Scenario 4 ($G_1(\tau_2) = G_1(\tau_1) + .30$) and Scenario 5 ($G_1(\tau_2) = G_1(\tau_1) + .10$). The final part of Exhibit IIIA is a pair of plots for each of $G_1(\tau_1) = .50, .65, .25$. The first plot in each pair is a plot of the reversal probability as a function of test reliability shown for each labeled value for Student 2 true percentile (e.g., for $G_1(\tau_1) = .50$ values of $G_1(\tau_2) = .60, .65, .70, .75, .80$). The second plot is reversal probability as a function of $G_1(\tau_2)$ for each labeled value of test reliability (.75, .80, .85, .90, .95).

Insert Exhibit IIIA here

Exhibit IIIA
Comparing Two Students: Reversal Probability

Scenario 1: Student 1 at 50th percentile, Student 2 at 75th percentile with perfect measurement: $G_1(\tau_1) = .50$, $G_1(\tau_2) = .75$.

Table IIIA1
Reversal probability as a function of test reliability.

rel	Pr{reversal}
0.75	0.204
0.775	0.188
0.8	0.17
0.825	0.15
0.85	0.128
0.875	0.104
0.9	0.0762
0.925	0.047
0.95	0.0188

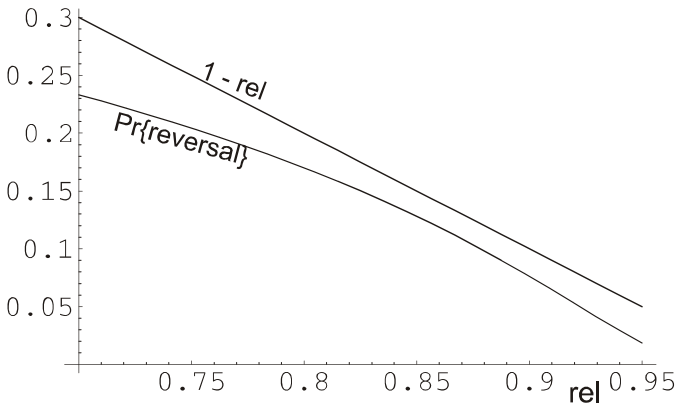


Figure IIIA1. Reversal probability as a function of test reliability; plot of 1-rel shown for comparison.

Exhibit IIIA continued

Scenario 2: Student 1 at 75th percentile, Student 2 at 90th percentile with perfect measurement: $G_1(\tau_1) = .75$, $G_1(\tau_2) = .90$.

Table IIA2
Reversal probability as a function of test reliability.

rel	Pr{reversal}
0.75	0.229
0.775	0.213
0.8	0.195
0.825	0.176
0.85	0.153
0.875	0.128
0.9	0.0989
0.925	0.0658
0.95	0.0307

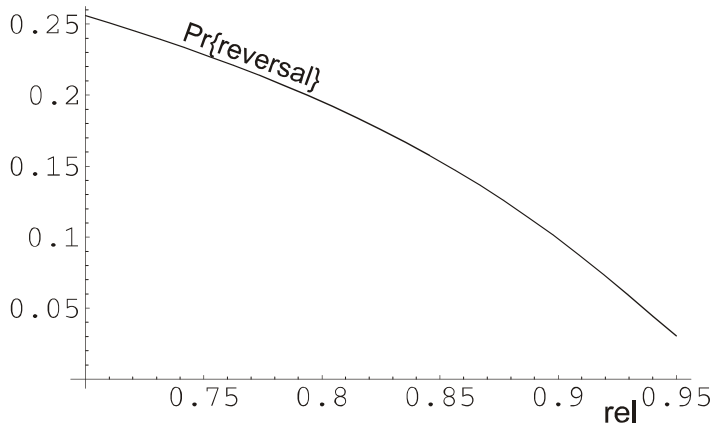


Figure IIIA2. Reversal probability as a function of test reliability.

Exhibit IIIA continued

Scenario 3: Student 2 is 20 percentile points greater than Student 1, with perfect measurement: $G_1(\tau_2) = G_1(\tau_1) + .20$.

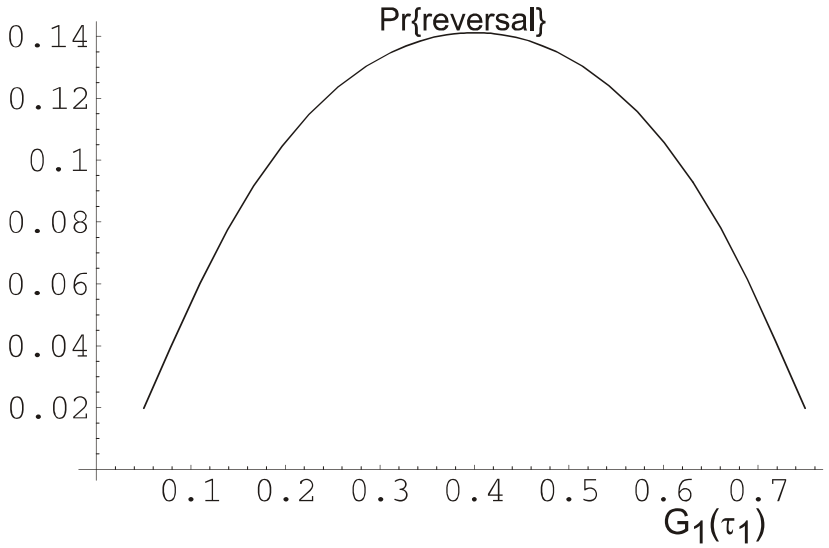


Figure IIIA3. Reversal probability for test reliability .90.

Table IIIA3. Reversal probability as a function of test reliability for Scenario 3.

	$G_1(\tau)$						
	.1	.2	.3	.4	.5	.6	.7
rel							
0.70	0.207	0.263	0.286	0.292	0.286	0.263	0.207
0.725	0.192	0.25	0.274	0.28	0.274	0.25	0.192
0.75	0.177	0.236	0.26	0.267	0.26	0.236	0.177
0.775	0.16	0.22	0.246	0.253	0.246	0.22	0.16
0.8	0.142	0.203	0.229	0.237	0.229	0.203	0.142
0.825	0.123	0.183	0.21	0.218	0.21	0.183	0.123
0.85	0.101	0.161	0.189	0.197	0.189	0.161	0.101
0.875	0.0783	0.136	0.163	0.172	0.163	0.136	0.0783
0.9	0.0541	0.106	0.133	0.141	0.133	0.106	0.0541
0.925	0.03	0.072	0.0964	0.104	0.0964	0.072	0.03
0.95	0.00981	0.0349	0.053	0.0592	0.053	0.0349	0.00981

Exhibit IIIA continued

Scenario 4: Student 2 is 30 percentile points greater than Student 1, with perfect measurement: $G_1(\tau_2) = G_1(\tau_1) + .30$.

Table IIIA4. Reversal probability as a function of test reliability for Scenario 4.

	$G_1(\tau)$					
	.1	.2	.3	.4	.5	.6
rel						
0.70	0.133	0.182	0.2	0.2	0.182	0.133
0.725	0.119	0.167	0.186	0.186	0.167	0.119
0.75	0.104	0.151	0.17	0.17	0.151	0.104
0.775	0.0886	0.135	0.154	0.154	0.135	0.0886
0.8	0.073	0.117	0.136	0.136	0.117	0.073
0.825	0.0572	0.0982	0.116	0.116	0.0982	0.0572
0.85	0.0418	0.0783	0.0952	0.0952	0.0783	0.0418
0.875	0.0272	0.0577	0.0728	0.0728	0.0577	0.0272
0.9	0.0146	0.0371	0.0495	0.0495	0.0371	0.0146
0.925	0.00534	0.0183	0.0267	0.0267	0.0183	0.00534
0.95	0.000764	0.00474	0.00826	0.00826	0.00474	0.000764

Scenario 5: Student 2 is 10 percentile points greater than Student 1, with perfect measurement: $G_1(\tau_2) = G_1(\tau_1) + .10$.

Table IIIA5. Reversal probability as a function of test reliability for Scenario 5.

	$G_1(\tau)$							
	.1	.2	.3	.4	.5	.6	.7	.8
rel								
0.70	0.317	0.366	0.385	0.392	0.392	0.385	0.366	0.317
0.725	0.307	0.358	0.378	0.386	0.386	0.378	0.358	0.307
0.75	0.295	0.349	0.37	0.378	0.378	0.37	0.349	0.295
0.775	0.282	0.339	0.361	0.37	0.37	0.361	0.339	0.282
0.8	0.267	0.327	0.351	0.36	0.36	0.351	0.327	0.267
0.825	0.25	0.313	0.339	0.349	0.349	0.339	0.313	0.25
0.85	0.229	0.297	0.324	0.335	0.335	0.324	0.297	0.229
0.875	0.205	0.276	0.306	0.318	0.318	0.306	0.276	0.205
0.9	0.175	0.25	0.283	0.295	0.295	0.283	0.25	0.175
0.925	0.137	0.215	0.25	0.265	0.265	0.25	0.215	0.137
0.95	0.0876	0.164	0.202	0.217	0.217	0.202	0.164	0.0876

Exhibit IIIA continued

Student 1 at 50th percentile, with perfect measurement: $G_1(\tau_1)=.5$.

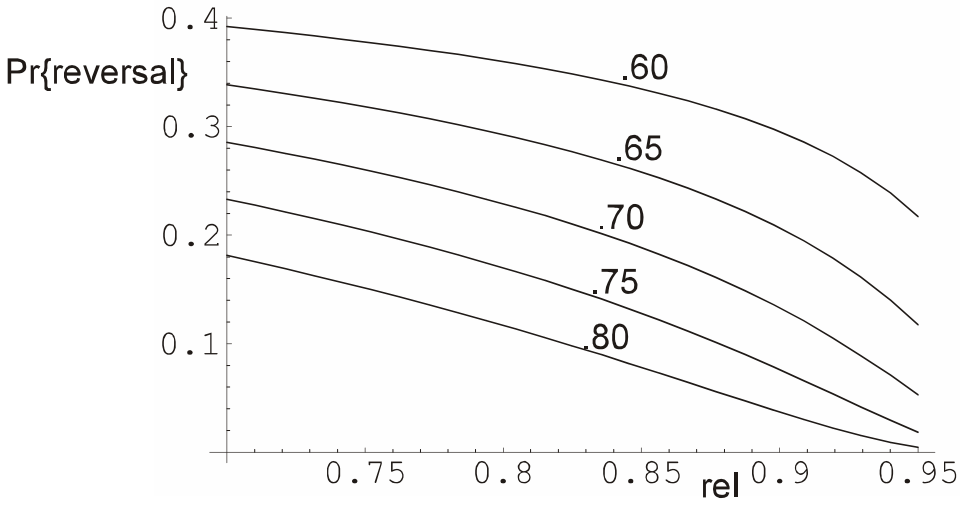


Figure IIIA4. Plot of reversal probability for each labeled value for $G_1(\tau_2)$ (Student 2 true percentile) as a function of test reliability.

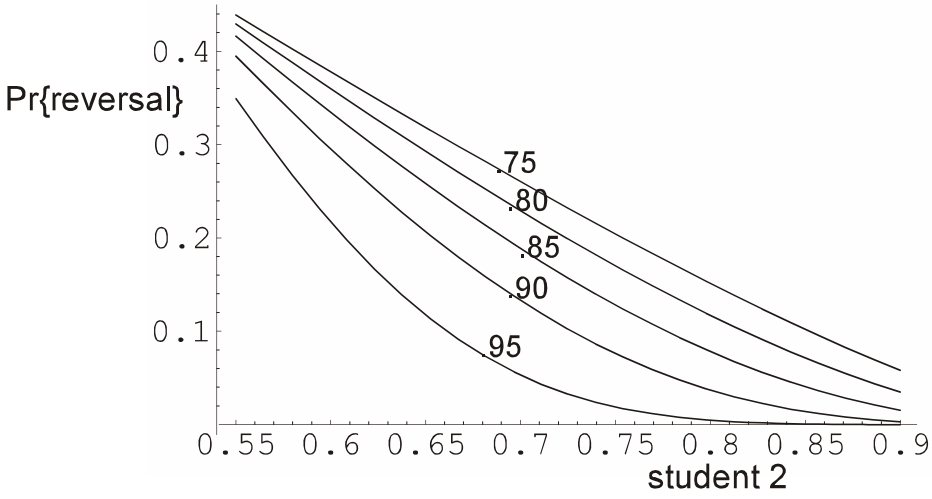


Figure IIIA5. Plot of reversal probability for each labeled value of test reliability as a function of $G_1(\tau_2)$ (Student 2 true percentile) .

Exhibit IIIA continued

Student 1 at 65th percentile with perfect measurement, $G_1(\tau_1)$=.65.

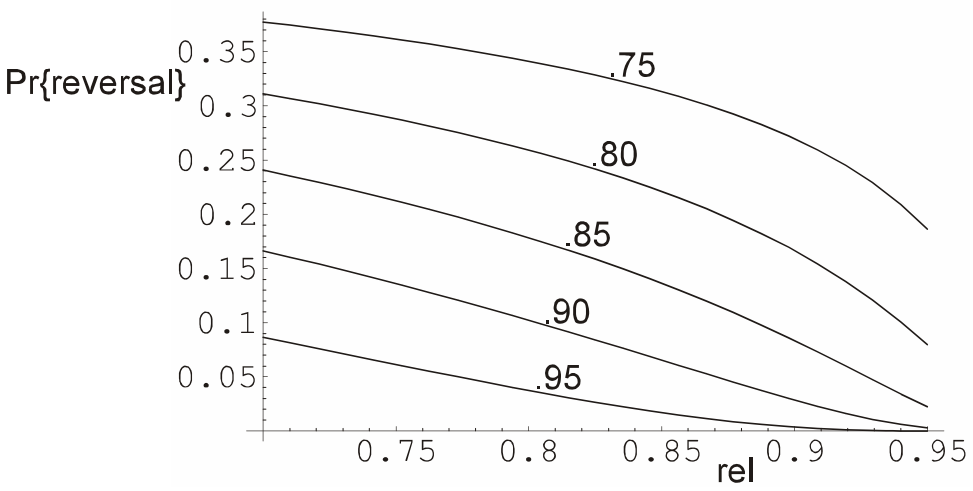


Figure IIIA6. Plot of reversal probability for each labeled value for $G_1(\tau_2)$ (Student 2 true percentile) as a function of test reliability.

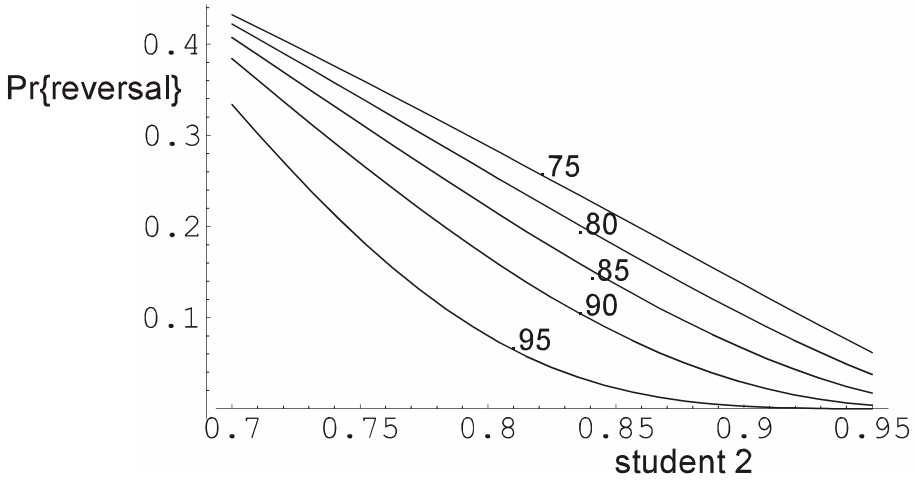


Figure IIIA7. Plot of reversal probability for each labeled value of test reliability as a function of $G_1(\tau_2)$ (Student 2 true percentile).

Exhibit IIIA continued

Student 1 at 25th percentile, with perfect measurement, $G_1(\tau_1)$=.25.

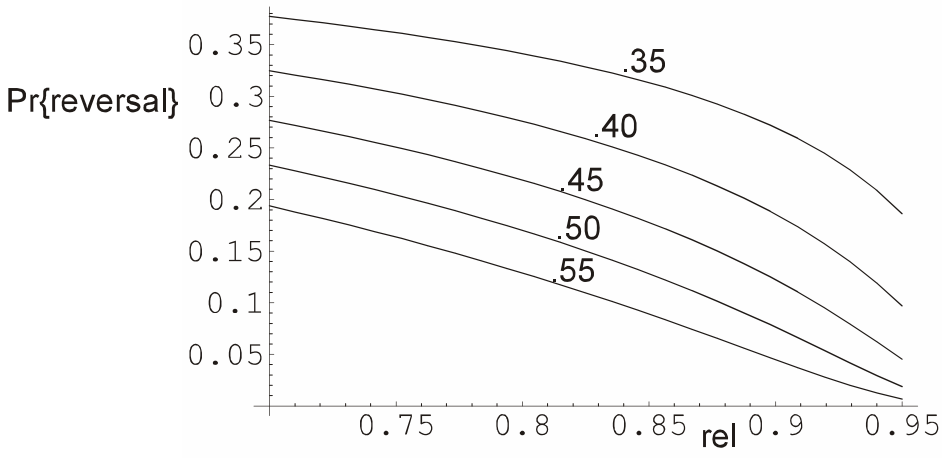


Figure IIIA8. Plot of reversal probability for each labeled value of $G_1(\tau_2)$ (Student 2 true percentile) as a function of test reliability.

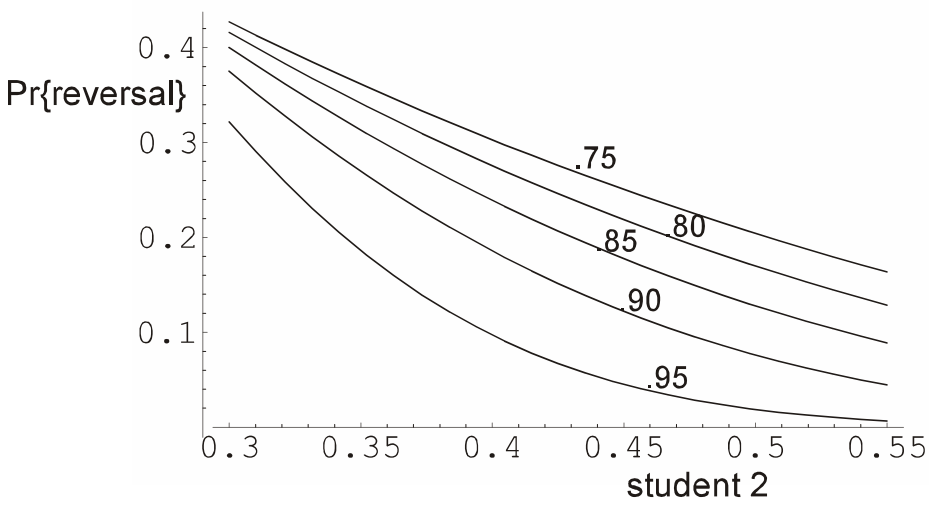


Figure IIIA9. Plot of reversal probability for each labeled value of test reliability as a function of $G_1(\tau_2)$ (Student 2 true percentile).

5.2 Difference Between Two Student's Percentile Scores

To extend the results on reversals, now examine $G(S_1) - G(S_2)$, the signed difference between the percentile rank scores for Student 1 and Student 2. Calculations in Exhibit IIIB present the quantity:

$$\text{compare2} = \Pr\{ G(S_1) - G(S_2) \le \text{bound} \mid G_1(\tau_1), G_1(\tau_2)\} , \tag{5.2}$$

the probability that the signed difference of the percentile ranks is less than or equal to the quantity "bound" (bound may be negative or positive) for two students with values of $G_1(\tau_1)$, $G_1(\tau_2)$. For bound set to 0, 1 – compare2 gives the reversal probability for $G_1(\tau_1) < G_1(\tau_2)$.

Computation of Compare2 Probability: Technical Details
The computation of the compare2 probability is implemented, in a manner similar to retest, using the following conditioning argument. For a Student 2 having a specified $G_1(\tau_2)$, condition on a draw of an s_2 from the S_2-distribution ($S_2 \mid \tau_2 \sim N[\, \tau_2, \sigma_N(1 - \text{rel})^{1/2} \,]$) and express that S_2-value in terms of its fractile of the S_2-distribution, ps_2 , to obtain:

$$\Pr\{G(S_1) - G(S_2) \le \text{bound} \mid ps_2 \} = \Pr\{ S_1 \le G^{-1}[G(S_2) + \text{bound}] \mid ps_2 \} =$$

$$\Phi\Big[\{\Phi^{-1}[\Phi[(1 - \text{rel})^{1/2}\, \Phi^{-1}[ps_2] + (\sqrt{\text{rel}})\, \Phi^{-1}[G_1(\tau_2)] \,] + \text{bound}] - (\sqrt{\text{rel}})\, \Phi^{-1}[G_1(\tau_1)] \}/(1 - \text{rel})^{1/2}\Big] \tag{5.3}$$

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

$$\text{compare2} = \int_0^1 \Big[\, \Phi[\{\Phi^{-1}[\Phi[(1 - \text{rel})^{1/2}\, \Phi^{-1}[ps_2] + (\sqrt{\text{rel}})\, \Phi^{-1}[G_1(\tau_2)] \,] + \text{bound} \,] - (\sqrt{\text{rel}})\, \Phi^{-1}[G_1(\tau_1)] \}/(1 - \text{rel})^{1/2}]\Big] d ps_2 \tag{5.4}$$

An alternative derivation is to condition instead on a draw of an S_1 through the fractile ps_1:

$$\Pr\{G(S_1) - G(S_2) \le \text{bound} \mid ps_1 \} = 1 - \Pr\{G(S_2) - G(S_1) \le -\text{bound} \mid ps_1 \} =$$
$$1 - \Pr\{S_2 \le G^{-1}[\, G(S_1) - \text{bound}] \mid ps_1 \} =$$
$$1 - \Big[\Phi[\{\Phi^{-1}[\Phi[(1 - \text{rel})^{1/2}\, \Phi^{-1}[ps_1] + (\sqrt{\text{rel}})\, \Phi^{-1}[G_1(\tau_1)] \,] - \text{bound} \,] -$$
$$(\sqrt{\text{rel}})\, \Phi^{-1}[G_1(\tau_2)] \}/(1 - \text{rel})^{1/2}]\Big] \tag{5.5}$$

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

$$\text{compare2} = \int_0^1 \Big[1 - \big(\Phi[\{\Phi^{-1}[\Phi[(1-\text{rel})^{\frac{1}{2}} \Phi^{-1}[ps_1] + (\sqrt{\text{rel}})\, \Phi^{-1}[G_1(\tau_1)]\,] - \text{bound}\,] - (\sqrt{\text{rel}})\, \Phi^{-1}[G_1(\tau_2)]\,\}/(1-\text{rel})^{\frac{1}{2}}]\big)\Big] d ps_1 \qquad (5.6)$$

Numerical illustrations. The three tables in Exhibit IIIB use the following specifications, respectively: $G_1(\tau_1) = .75$, $G_1(\tau_2) = .50$; $G_1(\tau_1) = .50$, $G_1(\tau_2) = .60$; $G_1(\tau_1) = .90$, $G_1(\tau_2) = .75$. For example, with $G_1(\tau_1) = .50$, $G_1(\tau_2) = .60$, test reliability .85 yields the compare2 result of a 12.5% chance (1 out of 8) that Student 1 (lower student on true level) obtains a test score at least 15 percentile points higher than Student 2.

Insert Exhibit IIIB here

Exhibit IIIB
Comparing Two Students: Difference of Percentile Ranks

Table IIIB1
$G_1(\tau_1) = .75$, $G_1(\tau_2) = .50$; Student 1 at 75th percentile, Student 2 at 50th percentile with perfect measurement.
$Pr\{ G(S_1) - G(S_2) \leq \text{bound} \mid G_1(\tau_1), G_1(\tau_2)\}$

	bound					
	-.15	-.10	-.05	0.0	0.05	0.10
rel						
0.70	0.101	0.137	0.181	0.233	0.293	0.36
0.725	0.0894	0.124	0.167	0.219	0.28	0.348
0.75	0.0778	0.111	0.153	0.204	0.265	0.335
0.775	0.0658	0.0967	0.137	0.188	0.249	0.32
0.8	0.0536	0.0819	0.12	0.17	0.232	0.304
0.825	0.0414	0.0666	0.102	0.15	0.211	0.285
0.85	0.0296	0.0509	0.0828	0.128	0.188	0.264
0.875	0.0188	0.0354	0.0624	0.104	0.161	0.238
0.9	0.00961	0.0208	0.0414	0.0762	0.13	0.206
0.925	0.00324	0.00884	0.0215	0.047	0.0924	0.165
0.95	0.000387	0.00169	0.0061	0.0188	0.049	0.11

	bound				
	.15	.20	.25	.30	.35
rel					
0.70	0.431	0.505	0.579	0.651	0.719
0.725	0.421	0.498	0.575	0.65	0.721
0.75	0.411	0.49	0.571	0.65	0.723
0.775	0.399	0.482	0.567	0.65	0.727
0.8	0.386	0.473	0.563	0.65	0.732
0.825	0.371	0.463	0.559	0.652	0.738
0.85	0.353	0.451	0.554	0.654	0.746
0.875	0.331	0.437	0.549	0.658	0.758
0.9	0.304	0.419	0.544	0.665	0.774
0.925	0.267	0.395	0.538	0.677	0.797
0.95	0.213	0.358	0.531	0.699	0.836

Table IIIB2
$G_1(\tau_1) = .50$, $G_1(\tau_2) = .60$; Student 1 at 50th percentile, Student 2 at 60th percentile with perfect measurement.
$Pr\{ G(S_1) - G(S_2) \leq \text{bound} \mid G_1(\tau_1), G_1(\tau_2)\}$ for Test Reliability .85

				bound						
-.25	-.20	-.15	-.10	-.05	0.0	.05	.10	.15	.20	.25
0.215	0.294	0.383	0.478	0.574	0.665	0.747	0.818	0.875	0.918	0.949

Exhibit IIIB continued

Table IIIB3
Student 1 at 90th percentile, Student 2 at 75th percentile with perfect measurement.
$\Pr\{ G(S_1) - G(S_2) \leq \text{bound} \mid G_1(\tau_1), G_1(\tau_2)\}$

	bound				
	-.10	-.05	0.0	0.05	0.10
rel					
0.70	0.125	0.181	0.256	0.346	0.444
0.725	0.112	0.168	0.243	0.336	0.437
0.75	0.0991	0.153	0.229	0.324	0.43
0.775	0.0853	0.138	0.213	0.311	0.422
0.8	0.0711	0.121	0.195	0.296	0.413
0.825	0.0565	0.103	0.176	0.278	0.402
0.85	0.042	0.0832	0.153	0.258	0.389
0.875	0.028	0.0627	0.128	0.233	0.373
0.9	0.0156	0.0416	0.0989	0.203	0.352
0.925	0.00599	0.0216	0.0658	0.163	0.324
0.95	0.000924	0.00616	0.0307	0.11	0.28

	bound				
	.15	.20	.25	.30	.35
rel					
0.70	0.54	0.63	0.71	0.78	0.838
0.725	0.538	0.633	0.716	0.788	0.847
0.75	0.536	0.636	0.724	0.797	0.857
0.775	0.534	0.64	0.732	0.808	0.868
0.8	0.532	0.644	0.741	0.821	0.881
0.825	0.53	0.65	0.753	0.835	0.896
0.85	0.528	0.658	0.768	0.852	0.912
0.875	0.526	0.668	0.786	0.872	0.93
0.9	0.523	0.682	0.809	0.896	0.949
0.925	0.52	0.703	0.841	0.926	0.97
0.95	0.516	0.738	0.887	0.961	0.989

6. Discussion/Remarks

An important role for a statistician is in the assessment of uncertainty. The implementation of accuracy in this paper can be seen as the flip-side of uncertainty in terms of statements like, How certain can I be that my reported percentile rank score is close to the target? The calculations of this paper use probability statements (such as probability that the observed percentile rank differs from true-rank by no more than the stated tolerance) to provide useful information on the accuracy of reported percentile rank scores (which are the scores most often reported to parents, media, etc.) The position of this paper in the presentation of results is that the Exhibits should be left to speak for themselves. The reader can be the judge of what is tolerable inaccuracy in a reported test score; this paper merely seeks to lay out some facts. However, one assertion is pretty clear: Conventional interpretations of test reliability coefficients do not well represent the accuracy of percentile rank scores (even though reliability coefficients are about all that is offered in most situations).

An interesting, and somewhat related, paper is the investigation of percentile rank scores in May and Nicewander (1994). Their calculations, based on an IRT formulation, employ more traditional criteria for accuracy; they seek (in part) to compare the reliability coefficient for the student observed-score with a reliability coefficient calculated for the corresponding percentile rank score (see their Table 2, p. 319). In the notation of this paper, the comparison might be constructed as $Var_p(\tau)/Var_p(Y)$ compared to $Var_p(G_1(\tau))/Var_p(G(Y))$ (where Var_p is a variance calculated over the population of persons), but such calculations won't be pursued at present. May and Nicewander find "that increased difficulty lowered the reliability of the PR [percentile rank] score at a substantially faster rate than the NR [number right, observed] score. In such cases an NR score of adequate reliability is accompanied by a PR that is virtually useless (in terms of reliability)" (p. 318). What these two papers have in common is a serious concern about the accuracy of the percentile rank score, and a finding that even for high reliability of the observed score, the percentile rank score may show disappointing properties. May and Nicewander conclude: "There exist situations in which one can reliably estimate the percentage of items known by examinees (using the percent-correct linear transformation of the NR score) but not the percentage of persons falling below a given score" (p. 325).

The limiting modifier “classical test theory” in the paper title is meant to communicate two points: what the present calculations are and a direction for further (similar) calculations. In this paper, the statement that the calculations are carried out for a classical test theory setting is merely intended to indicate calculations for Normally distributed, continuous scores, with constant error variance. Extension of these same calculations to empirical norms (other forms of G(Y)) and more complex measurement models (such as IRT scaling) where

measurement error variance depends on score value can be done in much the same manner. All that is needed are the specific forms of the norms distribution and the error variance. Such calculations will be the subject of further reports, for example, Rogosa (1999).

Another extension of this work, which will be reported in forthcoming Technical Reports, is to the accuracy of group summaries. A group of individuals could consist of a classroom of students, or larger collections of students such as a grade-level across a school or across a district. Especially in the California STAR program, accuracy properties of the group summary obtained from the National Percentile Rank of the student mean score (or median score) are of interest. Certainly, a group summary would be expected to have greater accuracy than an individual score; the question is, How much greater accuracy?

Author's Footnote

The basic material leading to the calculations on hit-rate and percentile discrepancy in Section 3.1 has some history. Those calculations grew out of my work in the Technical Study Group (TSG), Standards Curriculum and Assessment Division, California State Department of Education, with an initial presentation of that material to this group in September 1996, in regard to accuracy criteria for tests submitted for PTIP (Pupil Testing Incentive Program). Also, an early version of the test-retest consistency calculations in Section 4 was presented to the TSG group in February 1997, in the context of the California Comparability activities. Special acknowledgments go to the helpful comments and suggestions from William Schmidt and Richard Wolfe over the past three years. Furthermore, acknowledgment goes to Ross Green, CTB/McGraw-Hill, for independently proposing in April 1996 a version of the hit-rate criterion for a calculation like Table IE1 (for tolerance = 20). Moreover, this report benefitted from discussions with Joan Herman, CRESST, from review comments by Robert Mislevy of ETS, and from editorial assistance from CRESST staff.

References

ABC News (1998), *Primetime Live*, August 12, 1998.

Abram, N. (1996). *Measure Twice, Cut Once; Lessons From A Master Carpenter*. New York: Little Brown and Company.

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.

California Department of Education, *Proceedings of Comparability Symposium*, Burlingame, CA, October 1996.

Good Housekeeping Institute Report: Body-Fat Testers. *Good Housekeeping*, September 1998, p. 42.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

May, K., & Nicewander, W. A. (1994). Reliability and information functions for percentile ranks. *Journal of Educational Measurement*, *31*, 313-325.

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