Off Track: Problems with "On Track" Inferences in Empirical and Predictive Standard Setting

*Andrew Ho, Harvard Graduate School of Education*

October 24, 2012

**Abstract**

Predictive statements about future student performance can guide the selection of a performance standard. These predictive statements often take the following form: "A student with a current score $X$ has a $P$% chance of exceeding a future score $Y$." This paper uses straightforward regression relationships to clarify that these statements are ill-suited to support standard setting, as they confound the stringency of the selected cut score with the empirical relationship between $X$ and $Y$. A distinction is drawn between a cut score that *corresponds with* a future cut score and a cut score that *predicts* a future cut score. The latter is an indicator of the relationship between tests as much as the performance of students and is therefore subject to misinterpretation as a performance standard. This finding is unsurprising once stated but has considerable practical implications and does not appear to have been well described in the literature. Theoretical scenarios illustrate the policy-relevant consequences of this confounding, such as illusory stringency and leniency.

---

Recent educational reform efforts have focused attention on student readiness for college (U.S. Department of Education, 2010). These efforts have motivated "predictive standard setting": the selection of benchmarks or cut scores that indicate a probability of future success. Examples of college readiness benchmarks include those of ACT, that defines its benchmarks where students "have approximately a 50% chance of earning a B or better and approximately a 75% chance or better of earning a C or better in the corresponding college course or courses" (ACT, 2007, p. 24). The College Board defines its benchmarks as "the SAT score associated with a 65 percent probability of earning a first-year GPA of 2.67 (B-) or higher" (Wyatt, Kobrin, Wiley, Camara, & Proestler, 2011, p. 5). In this paper, I argue that these statements are accurate empirically but become misleading when used to set and describe standards for student performance.

An educational test score is always interpreted as an indicator of likely student performance in a broader domain. For predictive inferences about college readiness, validation

requires, in part, an investigation of the predictive relationship between test scores and college outcomes. When relationships are strong, the validation argument is strengthened, and when relationships are weak, predictive inferences must be tempered. Predictive standard setting seems to address this directly, by attaching empirically defensible predictive statements to particular regions of the test score scale.

Performance standards combine two concepts: They are 1) levels of performance that are expressed on 2) a well-defined scale. Justifying a level of performance without a well-defined scale is an example of the logical fallacy of "begging the question," where a part of the interpretive argument (appropriate performance level) is assumed to be the whole of the interpretive argument (appropriate interpretation; Kane, 2006). No matter how technically defensible a standard setting procedure may be, a logically necessary precondition is that the test score scale supports the desired inference (Haertel & Lorie, 2004).

As I will demonstrate, predictive standard setting confuses these two concepts, confounding the level of performance with the predictive utility of the scale. It is an effort to validate scale interpretations while simultaneously using the same evidence to determine the level of performance. As a result, the selected performance level is confounded with evidence for or against the validity of its underlying scale. A low standard may result in more "proficient" students not because performance in that region is associated with proficiency but rather because the test itself is a poor predictor of the outcome. In short, a low (or in some cases, high) standard may be evidence that the selection of a standard is itself unwarranted.

This confounding renders commonsense interpretations of performance levels impossible. Interpretations of stringency become tangled with the predictive utility of the test. Two standards that seem to differ in terms of stringency may instead differ due to differential predictive relationships with the future outcome. Disentangling stringency from predictive utility is not possible in practice, as performance level descriptors do not include the predictive evidence necessary to distinguish between the two concepts. Predictive standards support "on track" inferences only if they include the caveat, "to the extent that the test can predict the outcome." The counterintuitive relationships presented in this paper are evidence that these caveats are not natural to interpretations of performance standards and thus foster misinterpretation and misuse.

Familiar mathematical relationships from regression methods reveal the direction and degree of this confounding. Although the relationships require only basic statistical derivations, the implications for standard setting and "on track" inferences do not appear to be well described in the existing literature. I conclude that predictive analyses are poorly suited for selecting cut scores. As an alternative, I describe a three-pronged approach that uses an equipercentile or otherwise non-predictive standard setting procedure, supplementary statements about the predictive relationship between the two tests, and monitoring of the predictive relationship over time.

## Scenario 1: Simple Linear Regression

Let a "predictive standard setting approach" be one that uses regression-based analyses of longitudinal student data to support statements such as, "a student at or above a cut score, $X_c$, has a $P$% chance of exceeding a future cut score $Y_c$." I describe the variables, $P$ and $Y_c$, as associated with the "stringency" of $X_c$: The higher $P$ and $Y_c$ are, the higher and more "stringent" the selected cut score will be. This is intuitive and consistent with standard setting as an exercise in selecting a cut score on a well-defined scale.

The source of the confounding is the variable $\rho_{XY}$, the simple population correlation of observed scores from the two tests. This variable is not associated with stringency interpretations and yet has a direct impact on $X_c$ in predictive standard setting applications. In this first scenario, I ignore the probabilistic variable $P$ and focus on a predictive standard setting exercise that uses only simple linear regression.

One approach to separating stringency from correlation is to distinguish between a standard that *corresponds with* a future standard and a standard that *predicts* a future standard. One method that operationalizes correspondence—though by no means the only approach (Lewis & Haug, 2005)—is an equipercentile approach whose cut score, designated $X_c^{equi}$, establishes the same "passing" percentages for students on both $X$ and $Y$. One approach to operationalizing prediction is an Ordinary Least Squares (OLS) regression approach that results in the cut score, $X_c^{OLS}$. In comparing these two approaches, I present the equipercentile approach as an intuitive reference point, not an ideal. McClarty, Murphy, Keng, Turhan, and Tong (2012) helpfully compared cut scores set with these and other approaches, however they did not describe why or how these approaches were expected to differ theoretically.

A simple bivariate normal framework allows for interpretation of cut scores on a familiar "$z$-scale" of standard normal deviates. Under this model, both $X$ and $Y$ have standard normal distributions with a bivariate correlation, $\rho_{XY}$. The univariate distributions are identical, thus the equipercentile cut score is simply,

$$X_c^{equi} = Y_c. \quad (1)$$

Following simple linear regression relationships, the regression equation is $Y_c = \rho_{XY} X_c^{OLS}$, thus the cut score from the OLS approach arises from solving for $X_c^{OLS}$,

$$X_c^{OLS} = \frac{Y_c}{\rho_{XY}} \quad (2)$$

The relationship between the two cut scores follows:

$$X_c^{OLS} = \frac{X_c^{equi}}{\rho_{XY}}. \quad (3)$$

The difference between Equations 1 and 2 summarizes the distinction between a cut score determined solely by stringency (Equation 1) and a cut score that confounds stringency and association (Equation 2). The second cut score is skewed away from a "stringency-only" baseline whenever correlations are less than unity. Whenever future scores are negative (below average) on the $z$-scale, as they often are in practical applications (e.g., C grades or B- averages), Equation 3 shows that predictive standards will always be lower (more lenient) than equipercentile standards by a factor of $1/\rho_{XY}$.

Figure 1 displays the difference between these two approaches for a bivariate normal distribution with a correlation $\rho_{XY} = 0.6$. The equipercentile line is the diagonal (Equation 1), and the regression line is shown with slope $\rho_{XY}$ (Equation 2). A future cut score, $Y_c$, at the 20th and 80th percentiles (-.84 and +.84 respectively) leads to identical cut scores $X_c^{equi}$, but the regression-based cut scores $X_c^{OLS}$, relative to $X_c^{equi}$, diverge from the mean by the factor $1/\rho_{XY} \approx 1.67$.

This finding is elementary and straightforward from the standpoint of regression. In the absence of perfect correlation, regression will predict a future score closer to the future mean, $\bar{Y}$. Inverting to obtain a cut score on $X$ will result in a more divergent $X_c^{OLS}$. As an extreme but illustrative example, a completely uncorrelated predictor for a below-average $Y_c$ will have an infinitely lenient cut score, simply because the predicted future score will be $\bar{Y}$. Concluding from this that all students are "on track" is incorrect or at least incomplete. A more defensible conclusion is that all students are on track, as far as the predictor variable is capable of predicting (in this case, not at all). This crucial caveat turns the focus correctly back to the predictive relationship between the two variables and reveals the distortive influence of predictive association on standard setting. The proper response to this extreme example is to assert that the test scores on $X$ do not support interpretations about future performance on $Y$, thus no cut score is appropriate.

Equation 2 also offers an illustration of the difference between cut scores for two tests that have different correlations with future outcomes. In most longitudinal data, lower grades have lower correlations with future outcomes than higher grades. The ratio of their cut scores on the $z$-scale is the inverse of the ratio of their correlations. For example, if the correlation between a Grade 8 test and a high school exam is 0.7 and the correlation between a Grade 3 test and a high school exam is 0.5, then the cut score of the Grade 3 test will be 40% more lenient (or more stringent, if the future cut score is above the mean) than the cut score of the Grade 8 test, on the $z$-scale.

The top half of Figure 2 shows how cut scores will vary in their impact across grades, assuming that correlations between early grades and future outcomes will be lower than correlations between higher grades and future outcomes. The scale is expressed in terms of "impact data," the percentage of students expected to exceed the cut score $X_c^{OLS}$. For these standardized normal data, this is simply $\Phi(X_c^{OLS})$, where $\Phi$ is the standard normal cumulative distribution function. The top-right of Figure 2 shows impact data for a below-average future cut score at the 30$^{th}$ percentile. When correlations are higher, at 0.7 in Grade 10, around 77% of students are predicted to exceed the future cut score. If correlations drop to 0.3 for earlier grades, the percentage of proficient students in those grades will rise to 96%, simply because the predictive relationship between the earlier-grade test and the outcome is lower. If the future cut score is above the mean, as shown for a future cut score at the 70$^{th}$ percentile on the top-left of

Figure 2, earlier-grade tests will have much more stringent cut scores and lower percentages of proficient students.

It is important to note that the empirical reasoning that supports predictive standard setting is, strictly speaking, correct. Given the data, and to the extent that the regression equation matches the population regression equation for the target sample, performance at the cut score $X_c^{OLS}$ really does predict a future performance of $Y_c$. The cut score $X_c^{equi}$ is inferior as a predictor of this future performance. However, the proper criterion by which to evaluate any cut score $X_c$ is not by predictive accuracy but by the accuracy of the inferences that the cut score supports. These interpretations, following performance level descriptors, primarily involve a level of student performance and does not include a conditional inference about the predictive utility of the test. By confounding this performance level with predictive utility, predictive standard setting ends up answering two questions, about stringency and prediction, with one answer, and ultimately answers neither question accurately.

**Scenario 2: Probabilistic Regression**

Direct prediction of a future score can be extended to probabilistic prediction of a future score. In practice, this probabilistic interpretation is often supported by logistic regression models (ACT, 2007; Allen & Sconing, 2005; Wyatt, Kobrin, Wiley, Camara, & Proestler, 2011), where the outcome variable is dichotomized by the future cut score, $Y_c$, and a linear model is used for the log-odds of exceeding the cut score. A close cousin to the logistic model is the probit model, which results in effectively similar estimates but requires more complex estimation procedures. An attractive alternative to both of these is quantile regression (Koenker, 2005), which may be theoretically preferable to avoid the loss of information, power, and efficiency that comes with dichotomization of the outcome variable (Cox, 1957; Fedorov, Mannino, & Zhang; 2009; Ragland, 1992).

For notational convenience, I derive probabilistic predictive standard setting relationships for the probit model under bivariate normality. This generalizes directly to the quantile regression model under bivariate normality, and it is a close approximation to the logistic model to the extent that the logit function can approximate the probit function. The derivation follows from regression relationships that are well understood across disciplines and can even be found in justifications for the normal ogive model in Item Response Theory (e.g., Lord, 1980).

The conditional probability, $P$, of exceeding a future target or threshold $Y_c$, given an initial score $X$ that takes the value $X = x$, can be modeled with a normal ogive, Φ, with a "slope" parameter of $\beta_1$ and an "intercept" parameter of $\beta_0$.

$$\mathrm{P}(\mathrm{Y} > \mathrm{Y_c}|X = x) = \Phi(\beta_0 + \beta_1 x)$$

When $X$ and $Y$ are distributed as a bivariate normal distribution, the slope and intercept can be expressed in terms of the target cut score, $Y_c$, and the correlation, $\rho_{XY}$:

$$\beta_1 = \frac{\rho_{XY}}{\sqrt{1 - \rho_{XY}^2}}$$

$$\beta_0 = \frac{-Y_c}{\sqrt{1 - \rho_{XY}^2}}$$

Substituting and solving for $X_c = x$ gives the target expression,

$$X_c^{prob} = \frac{\Phi^{-1}(P)\sqrt{1 - \rho_{XY}^2} + Y_c}{\rho_{XY}}. \qquad (4)$$

Approximation of slopes and intercepts from a logistic regression model could be obtained by simply multiplying them by 1.702 (Camilli, 1994). However, the differential fit between a logistic model and a probit model to bivariate normal data is negligible and makes little difference to this presentation. The key terms are again the stringency-related variables, $P$ and $Y_c$, which together have intuitive and predictable effects on the probabilistic cut score $X_c^{prob}$. It is the correlation, $\rho_{XY}$, that again represents a confounding factor that confuses stringency with predictive utility. Note that when the probability, $P$, is 50%, Equation 4 reduces to Equation 2, as expected.

Figure 3 displays some of the implications of Equation 4. The height of the bars represent the stringency of cut scores on the $z$-scale. The left-most black bar in each grouping represents the intuitive reference point of the equipercentile cut score, indicating a "stringency-

only" baseline. To mimic operational uses, the target score, $Y_c$, is restricted to below-average scores and expressed as a percentile, and $P$ is greater than 50% (e.g., a 65% chance of a B-average). The top half of Figure 2 shows cut scores when $\rho_{XY} = 0.45$, and the bottom half of Figure 2 shows cut scores when $\rho_{XY} = 0.7$. This range of correlations covers many intergrade correlations seen in practice (e.g., Claessens, Duncan, & Engel, 2009) and sometimes exceeds correlations between achievement tests and postsecondary outcomes (Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008; Noble & Sawyer, 1987).

The "50% chance" bars are the cut scores set by the median regression line. The median line is equal to the OLS regression line under bivariate normality, thus the results follow from Equation 3, where the cut score is always more lenient until the future cut score reaches the mean. As expected, the lower correlation in the top figure leads to a greater discrepancy between the equipercentile cut score, $X_c^{equi}$ (in black), and the OLS-based cut score, $X_c^{OLS}$, that is equal to the median regression-based cut score in dark gray. Within each figure, increasing $P$ raises the stringency of cut scores (each successive bar is more positive), and increasing the future cut score raises the stringency of cut scores (each successive set of bars is more positive). These relationships are, I argue, intuitive reflections of standards-based interpretations. The dependence on the correlation, that is, the difference between the top and bottom figures, is not.

Returning to Figure 2, the bottom half of the figure shows the implications of Equation 4 for a 2/3 ≈ 67% chance of reaching a future outcome at the 70$^{th}$ or 30$^{th}$ percentiles. The higher probability of reaching a high future standard, in the bottom left of the figure, results in extremely low percentages of students exceeding the cut score, as expected from the combination of stringency, low association, and an above-average future cut score.

The bottom right of Figure 2 shows that, for this particular combination of $P = 67\%$ and $Y_c = \Phi^{-1}(30\%)$, the cut score $X_c^{prob}$is quite similar over correlations between 0.3 and 0.7. This follows directly from the multidimensional surface that Equation 4 describes, where the marginal slope of $X_c^{prob}$ on $\rho_{XY}$is shallow for general statements of the form, "moderately high probabilities of reaching moderately low cut scores." This finding may be practically reassuring but does not weaken any of the theoretical arguments in this paper. First, the dependence of $X_c^{prob}$on $\rho_{XY}$ is still nonzero, and Equation 4 can be used to show the exact difference in correlations that predict a consequential difference in impact data, however "consequential" is

defined. Second, more importantly, it is just as easy to describe this result as one that demonstrates that prediction does not matter for predictive standard setting. This defeats the whole purpose of the exercise. If the motivation is to select a probabilistic statement that minimizes the impact of prediction, then describing the procedure as one driven by prediction is inaccurate. The next scenario expands on this argument.

**Scenario 3: When Probabilistic Regression Leads to Equipercentile Cut Scores**

There is always a combination of $P$, $Y_c$, and $\rho_{XY}$ where $X_c^{prob} = X_c^{equi}$. For example, as Figure 2 shows, the two are close when $\rho_{XY} = 0.45$ and the cut score is defined by a 70% chance of reaching the 20th percentile. If an equipercentile cut score represents a stringency-based inference, uncontaminated by predictive statements, then this observation allows for the layering of predictive statements onto stringency-driven cut scores. This can be formalized by solving for the probabilistic statements that result in equipercentile cut scores. To obtain a cut score equivalent to an equipercentile cut score, $X_c^{equi}$, for any given correlation, $\rho_{XY,}$ simply define a probability, $P^{equi}$, of reaching a future score $Y_c^{equi}$, as follows,

$$Y_c^{equi} = -\Phi^{-1}\left(P^{equi}\right)\sqrt{\frac{1+\rho_{XY}}{1-\rho_{XY}}}. \qquad (5)$$

Graphing these for correlations of 0.45 and 0.7 gives the curves shown in Figure 4. This reveals that, for any given correlation, a range of predictive statements support an equipercentile standard. When the correlation is 0.45, an equipercentile standard could be described as an 85% chance of reaching the 5th percentile or a 70% chance of reaching the 20th percentile (as noted in the previous paragraph). If the correlation is 0.7, an equipercentile standard could be justified by a 70% chance of reaching the 10th percentile. One could always justify with probabilistic statements what has been determined by equipercentile or other methods. This paper argues that probabilistic statements should not justify standards alone, as standards are interpreted primarily in terms of stringency. Predictive statements may enhance, but neither guide nor determine, selected cut scores.

**Recommendations**

Following Haertel and Lorie (2004), the necessarily primary goal should be gathering sufficient evidence that any cut score on $X$ can support inferences about $Y$. Predictive analyses should provide some of this evidence. Providing that sufficient evidence exists, a standard setting exercise should be conducted that does not include regression-determined guidance. This may be supported in part by equipercentile benchmarks but need not be restricted to equipercentile approaches. Finally, the chosen cut score may be enhanced with predictive statements, although I recommend extreme caution in this regard. These predictive statements threaten to skew interpretations back to a confounding of stringency and prediction that may lead to future confusion.

As an example, consider a standard setting endeavor that, following the above recommendations, results in a cut score that is essentially equipercentile in nature. If the relationship is bivariate normal and the correlation is 0.45, we might attach a probabilistic statement to this cut score following Figure 3: that performance at a cut score at the 20$^{th}$ percentile predicts a 70% chance of exceeding the 20$^{th}$ percentile in the future. Let us then assume that the same equipercentile approach is applied to a test from a more proximal grade with a correlation of 0.7. As Figure 4 shows, the probability of reaching the 20$^{th}$ percentile is now 63%, because the higher correlation increases the likelihood that a 20$^{th}$ percentile student remains distal from the mean instead of "regressing" towards it.

The difference between these the 63% and the 70% chances is explained by regression but is counterintuitive and at odds with the primary inference of stringency on the scale. The "70% chance" reflects a weaker argument for the selection of any cut score, not evidence that the cut score should be higher or lower. The best use of regression methods is therefore prior assessment of the predictive utility of the scale, before standards are set, rather than post-hoc rationalization of standards with predictive statements. If correlations are too low, then no standard warrants an indication of future performance. If they are sufficiently high, then a standard should be set that reflects stringency, without confounding the cut score with the very correlational evidence that supported the standard setting procedure.

Figure 1. Contrasting equipercentile and regression-based standard setting procedures for high and low future cut scores when $\rho_{XY} = 0.6$.

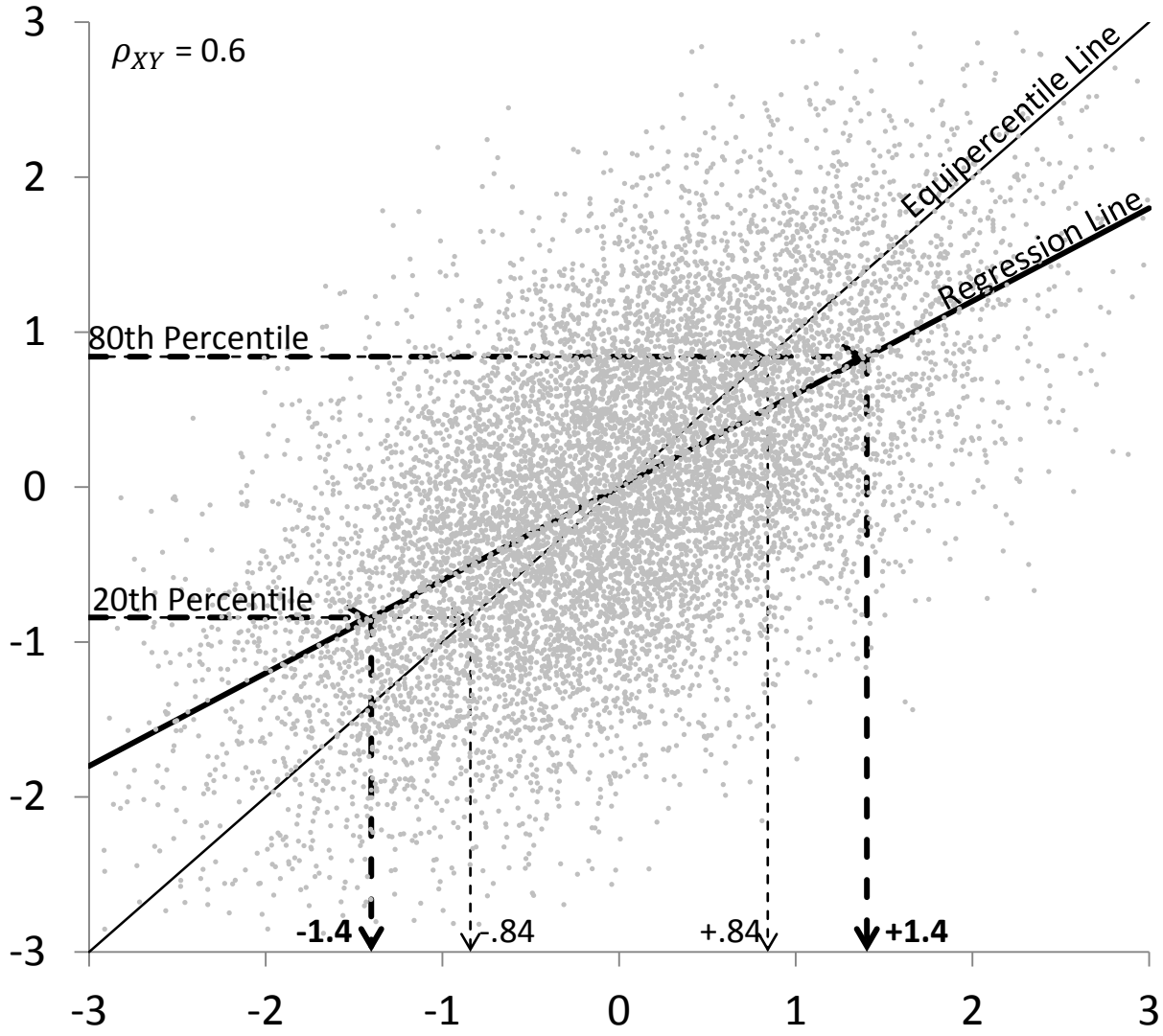


Figure 2. Across-grade impact data in terms of percentage of proficient students, assuming lower grades have lower correlations with future outcomes as shown.

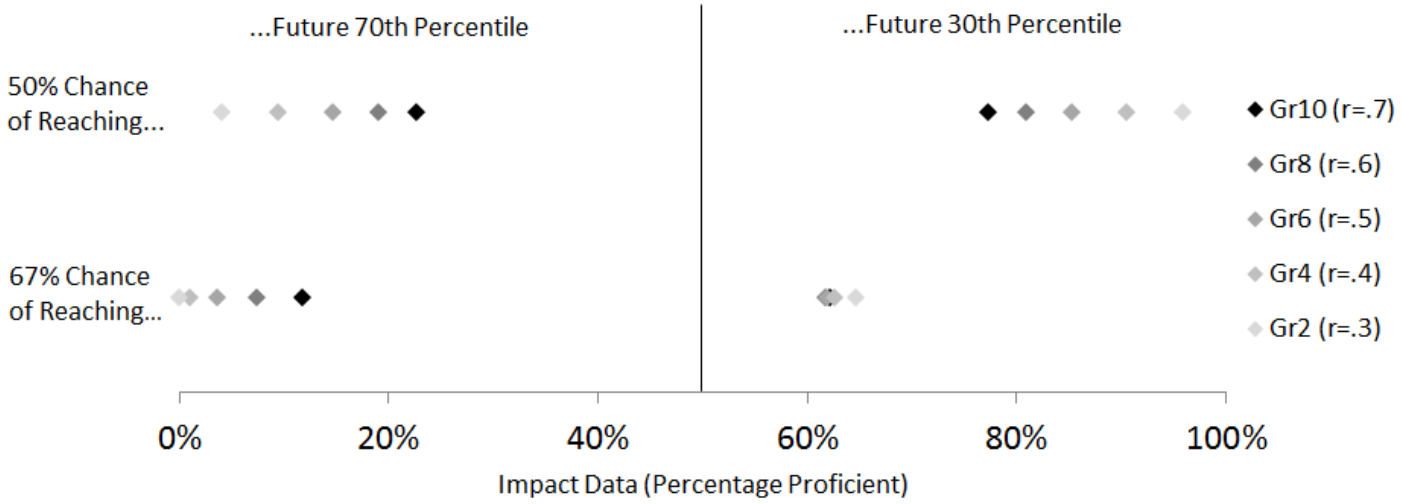
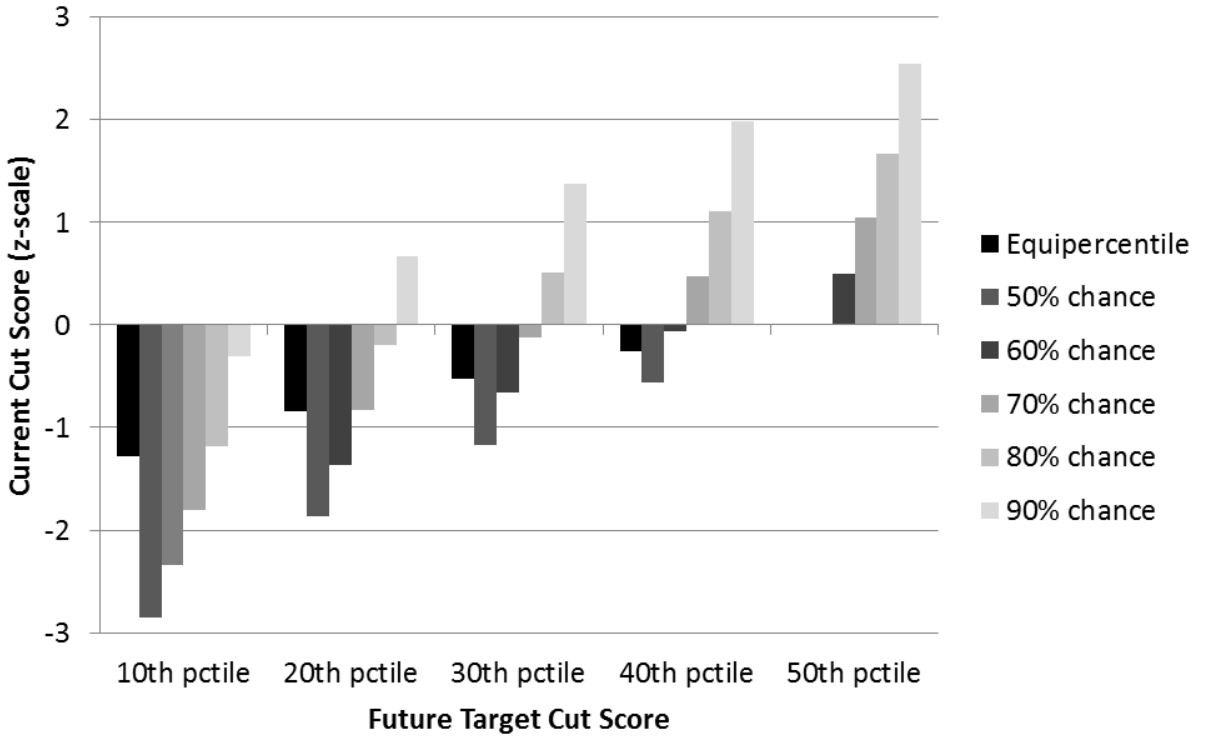


Figure 3. Contrasting equipercentile cut scores and probit regression cut scores for correlations of 0.45 and 0.7.



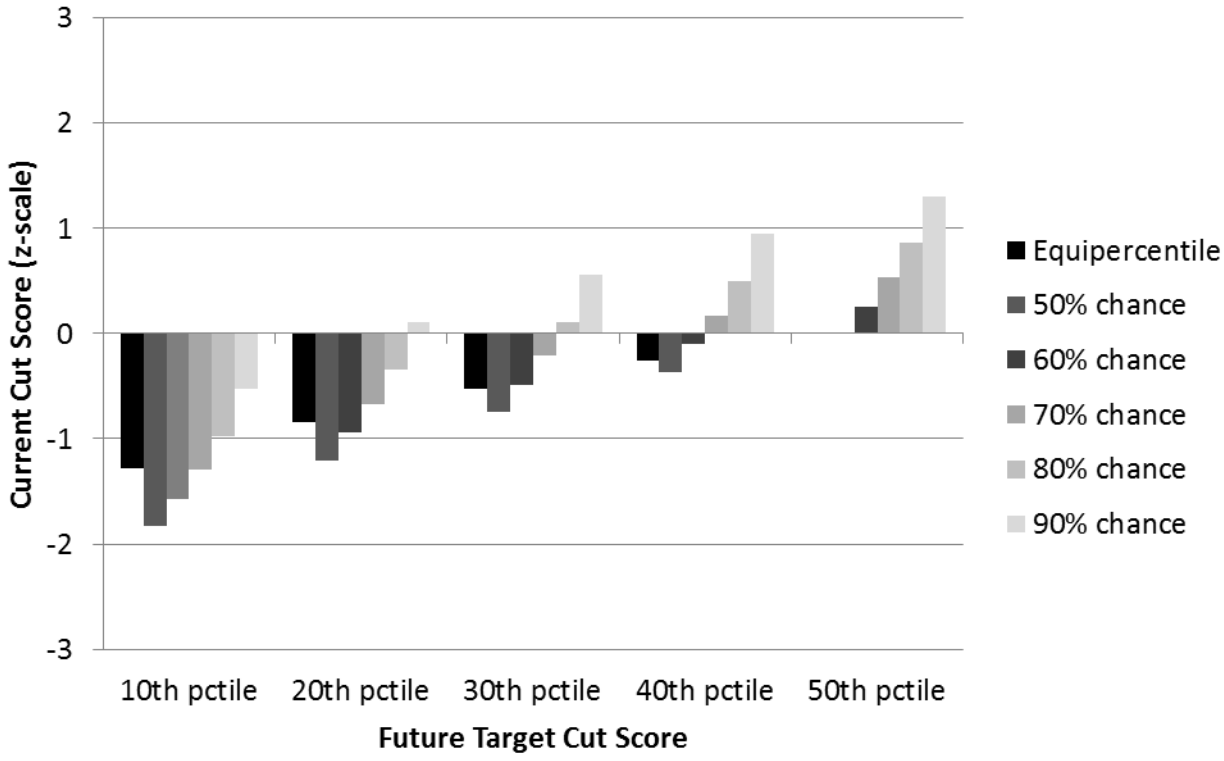
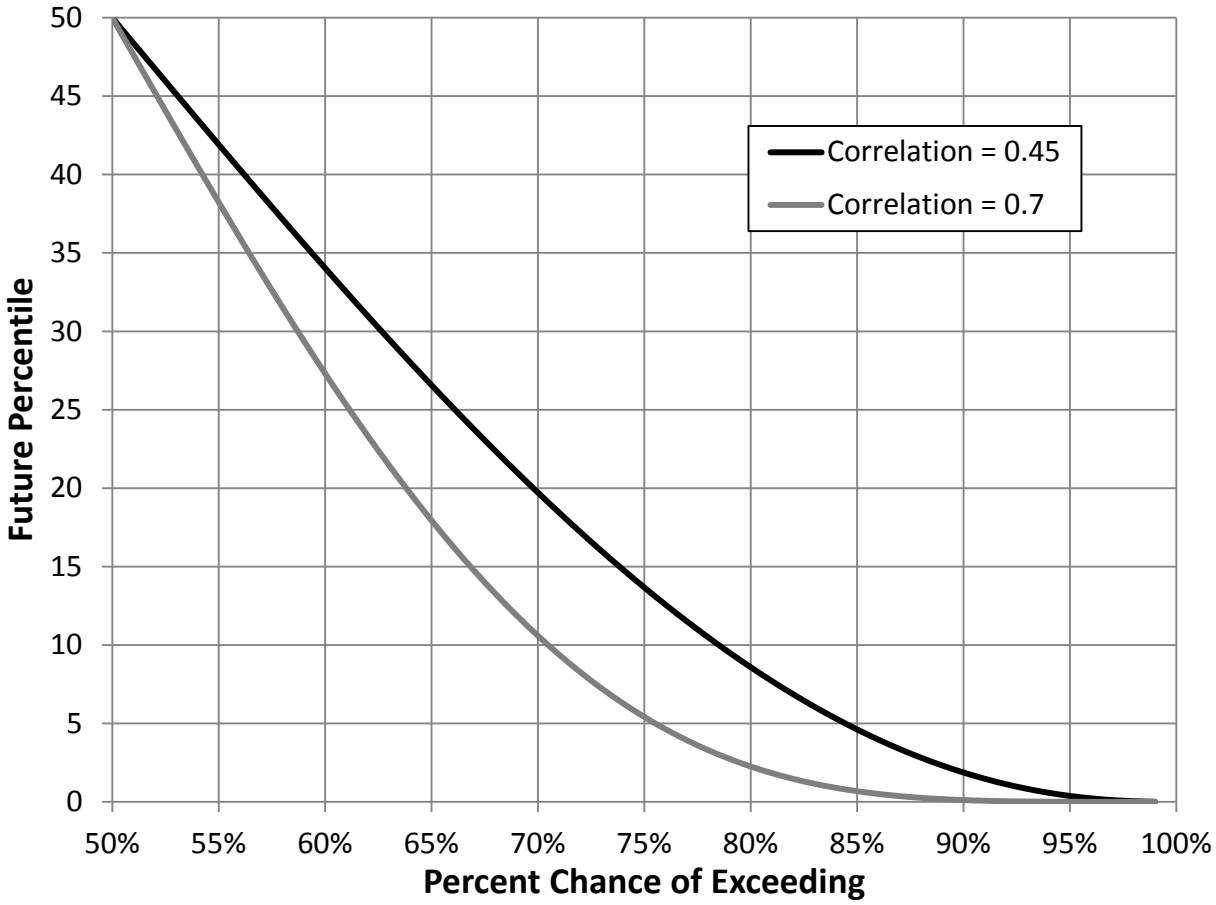



Figure 4. Predictive statements associated with equipercentile cut scores for correlations of 0.45 and 0.7.



## References


ACT (2007). The ACT Technical Manual. Iowa City, IA.

Allen, J. & Sconing, J. (2005). *Using ACT Assessment Scores to Set Benchmarks for College Readiness* (ACT Research Report Series 2005-3). Retrieved from http://www.act.org/research/researchers/reports/pdf/ACT_RR2005-3.pdf

Camilli, G. (1994). Teacher's corner: Origin of the scaling constant $d = 1.7$ in Item Response Theory. *Journal of Educational and Behavioral Statistics, 19,* 293-295.

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review, 4*, 415-427.

Cox, D. R. (1957). Note on grouping. *Journal of the American Statistical Association, 52,* 543-547.

Federov, V., Mannino, F., & Zhang, R. (2009). Consequences of dichotomization. *Pharmaceutical Statistics, 8,* 50-61.

Haertel, E. H., & Lorie, W. A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives, 2*(2), 61–103.

Kobrin, J. L, Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). Validity of the SAT for predicting first-year college grade point average. College Board Research Report #2008-5.

Koenker, R. (2005). *Quantile regression.* New York, NY: Cambridge University Press.

Lewis, D. M., & Haug, C. A. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education, 18,* 11-34.

McClarty, K. L., Murphy, D., Keng, L., Turhan, A., Tong, Y. (2012). Putting ducks in a row: Methods for empirical alignment of performance standards. Paper presented at the 2012 Annual Meeting of the National Council on Measurement in Education. Vancouver, Canada.

Noble, J. P., & Sawyer, R. (1987). Predicting grades in specific college freshman courses from ACT test scores and self-reported high school grades. ACT Research Report Series #87-20.

Ragland, D. R. (1992). Dichotomizing continuous outcome variables: Dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology, 3,* 434-440.

U.S. Department of Education (2010, March). A blueprint for reform: The reauthorization of the Elementary and Secondary Education Act. Office of Planning, Evaluation, and Policy Development. Washington, DC.

Wyatt, J., Kobrin, J., Wiley, A., Camara, W. J., & Proestler, N. (2011). SAT benchmarks: Development of a college readiness benchmark and its relationship to secondary and postsecondary school performance. College Board Research Report #2011-5.