Research Report

Providing a Context for Interpreting Predictions of Job Performance

ETS RR–18-38

Neil J. Dorans

December 2018

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Providing a Context for Interpreting Predictions of Job Performance

Neil J. Dorans

Educational Testing Service, Princeton, NJ

A distinction is made between scores as measures of a construct and predictions of a criterion or outcome variable. The interpretation attached to predictions of criteria, such as job performance or college grade point average (GPA), differs from that attached to scores that are measures of a construct, such as reading proficiency or knowledge about a domain (e.g., physics). In contrast to a score whose meaning is tied to the fidelity with which it measures a construct, the meaning of a prediction is tied to the statistical relationship of its scale to that of the criterion. A contrived example that might be seen in an employment setting is used to illustrate potential misinterpretations associated with making the predicted score distribution look like the criterion score distribution it is predicting.

Keywords Predictions; criterion; accuracy; scores; construct; scale alignment

doi:10.1002/ets2.12227

Test scores are often used to predict levels of a criterion or outcome variable. A classic example is the use of test scores in conjunction with high school grades to predict academic performance in college or university. These predictions are not scores in the traditional sense of a score as a measure of an attribute or construct. The first section explains the distinction between a score as a measure of a construct and as a prediction of an external criterion. The meaning of a prediction is tied to its statistical relationship with the criterion. That relationship is described in different ways by a scale-dependent standard error of prediction and by a scale-invariant correlation coefficient. When predictions are placed on a different scale via a linear transformation the correlation with the criterion score is unaffected. When predictions are transformed away from the metric of the criterion score however the average standard error of prediction increases unless the criterion score is transformed in the same manner as the predicted score. The consequences of rescaling predictions on the standard error of prediction are illustrated in the second section with a hypothetical example. The last section offers suggestions for reporting predictions.

The Distinction Between Scores and Predictions

Scores from tests used for academic admissions, such as the *SAT*®, ACT, and *GRE*® General Test, are reported on score scales. Each of these score scales has one universal meaning that is shared by many score users, including test takers, and a multitude of local meanings that are tied to specific local uses of the test score. Variation in institutional selectivity may lead different institutions to attend to different ranges of SAT or ACT scores. The score scale provides a framework for the interpretation of scores that is invariant across intended uses. A reported score of a given number for the SAT (or GRE or ACT) has the same universal meaning across different editions of that test that are placed on the same scale. The universal meaning of many of these scores is tied to the construct, such as proficiency in mathematics, reading, or writing, that they purport to measure and often to a reference population of test takers. Test takers are the user group that makes most legitimate use of the universal meaning of the scale. In the case of composite scores such as the ACT composite, which is the simple sum of equated English, math, reading, and science scores divided by 4, the universal meaning of the score scale is not tied directly to a clearly defined construct. The combination rule, however, is the same across all editions of the test, and this consistency provides some meaning, albeit a complex one from a content perspective, to the composite. As a result, the ACT composite derives some meaning from the consistent manner in which the four components are treated.[1]

Corresponding author: N. J. Dorans, E-mail: ndorans@ets.org

Scale Alignment Versus Prediction

Holland and Dorans (2006) made a distinction between different classes of score linkages. One major class is scale alignment. Scale alignment usually operates on numbers that are considered measures of some construct. All scale aligning procedures map two score distributions onto a common scale in some population of test takers so that the aligned scores represent the same relative position in that population. Typically this involves transforming scores from the metric of one test to that of another via linear linking (same standing with respect to standard deviation units from the average score) or equipercentile linking (same standing with respect to percentile rank). A variety of data collection designs and data analysis procedures can be used to align scales. Features of testing situations that affect the type of inference that is attached to the score can be divided into the following categories: test content, conditions of measurement, and examinee population. As described in Dorans, Pommerich, and Holland (2007), different types of scale alignments that vary along these and other dimensions include equating, linking tests in transition, concordance, and vertical scaling, among others. Holland and Dorans (2006) and others use the term *score equating* to refer to the aligning of score distributions from tests that produce scores that can be used interchangeably. Equated scores must measure the same construct, have the same reliability, and exhibit a linking relationship that is invariant across different subpopulations.

Another class of score linkages discussed by Holland and Dorans (2006) is *prediction*. Prediction is different from scale alignment or scaling. Whereas scale alignment matches score distributions, the goal of prediction is to predict an examinee's standing on an outcome variable from other information about that examinee, with minimal error of prediction. The information used for prediction might include a score on a test or scores from several other tests, and it might also include demographic or other information.

The goal is for the prediction to be close to the criterion being predicted. A good prediction would have the property of being calibrated, in the sense that the expected value of the criterion conditional on the prediction equals the prediction. For example, a prediction of undergraduate grade point average (UGPA) is calibrated or unbiased if among all people predicted to have a UGPA of say, 3.2, the expected value of their UGPA is 3.2.

Typically there are multiple predictors that are intentionally designed to measure different constructs. For example, UGPA is often predicted from high school grades, test scores that measure different skills or proficiencies, and other relevant information. Contrast this prediction with a score defined by the items on a test that are specifically designed to measure a single construct. All these items are included on the test because they are thought to measure the same construct. The predictors of UGPA, on the other hand, are included because they measure different constructs or variables that are believed to help predict academic performance in college.

There is always an asymmetry between what is predicted and what is being used to make the prediction. This asymmetry eliminates prediction as a form of scale aligning that strives to match score distributions between two variables (Holland & Dorans, 2006).

Reporting Scores Versus Contextualizing Predictions

In educational settings, the measurement of a score that measures a particular construct and the reporting of these scores on an established score scale are of primary interest. The choice of scale on which to report scores is one of a testing program's most fundamental and critical decisions. The usefulness of a score scale depends on how well it supports the inferences attached to its scores, how well it facilitates meaningful interpretations, and how well it prevents misinterpretations. In educational settings, validation tends to focus on validating the test as a measure of a construct, as documented extensively in Brennan (2006). Reported scores on a test often purport to measure a specific construct.

In contrast, the focus in employment settings is often on prediction of a criterion rather than measurement of a construct. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) is explicit about this emphasis on prediction. The Employment Testing section in Chapter 11, "Workplace Testing and Credentialing," states the following:

> The fundamental inference to be drawn from test scores in most applications of testing in employment settings is one of prediction: The test user wishes to make an inference from test results to some future job behavior or job outcome. Even when the validation strategy used does not involve empirical predictor-criterion linkages, as in the case of validity evidence based on test content, there is an implied criterion. Thus, although different strategies for

> gathering evidence may be used, the inference to be supported is that scores on the test can be used to predict subsequent job behavior. The validation process in employment settings involves the gathering and evaluation of evidence relevant to sustaining or challenging this inference. (American Educational Research Association et al., 2014, pp. 171–172)

A predicted criterion score does not typically measure the same construct as the criterion.[2] For example, a linear combination of ACT scores and high school grades that minimizes the mean squared error of prediction of college grades is not a measure of college grades in the way that scores from two versions of the ACT Math test measure the same construct. Likewise, a composite of personal characteristics that purports to predict success in some setting is not a measure of success in that setting.

Instead, the meaning of a predicted score is derived from the fact that it is the combination of a set of scores with the smallest average squared error of prediction of a particular set of criterion scores.[3] This criterion-specific meaning changes with each criterion and with each set of predictors for any particular criterion. Take for example the prediction of UGPA. It might be predicted from high school GPA (HSGPA). In the sample in which the prediction equation is determined, the predicted UGPA would represent the part of UGPA that is perfectly related to HSGPA.[4] The remainder or residual, which can be either positive or negative in value, is uncorrelated with the predicted part. If we add test scores such as the SAT (or ACT) scores to the battery of predictors, then the predicted UGPA is that part of UGPA that is perfectly related in that same sample to a linear combination of HSGPA and SAT (or ACT) scores. This part of UGPA predicted from both HSGPA and test scores is different from the prediction obtained only with HSGPA or from that obtained only with SAT (or ACT) scores. The meaning of each of these two predictions (and predictions in general) is statistical, varies as the composition of the set of predictors varies, and is not the same as the meaning of the criterion. In sum, whereas scores are measures of a construct reported on an established score scale with a shared universal meaning as well as situation-specific meanings, predictions are not.

An Example

Placing Predicted Criterion Scores on the Criterion Score Scale of 1 to 7

Suppose a testing process produces estimates of standing on each of several dimensions. For example, suppose there are three cognitive measures, reading, math and writing, plus measures of five personality traits, such as the Big Five (Goldberg, 1993). Without loss of generality, assume that the criterion score/rating, Y, is transformed to have a mean of 4 and standard deviation of 1. The regression of these criterion scores onto the eight dimensions produces a composite with a multiple correlation of R.

Let this composite of the predictor scores, $\mathrm{X_r}$, which has a mean of 4 and a standard deviation equal to R, be transformed to $\mathrm{X_s}$, which has a mean of 4 and a standard deviation of 1 via $\mathrm{X_s} = 4 + (\mathrm{X_r} - 4)/R$. Both Y and $\mathrm{X_s}$ are on scales with midpoints of 4 and with most values falling between 1 and 7. In contrast, $\mathrm{X_r}$ has a mean of 4 and a standard deviation equal to R, with most values falling between $4 - 3R$ and $4 + 3R$.

On what scale should we report the predicted criterion scores? An obvious choice is the criterion score scale, Y. What does it mean to report predicted scores on the criterion scale? It means using $\mathrm{X_r}$ as an estimate of Y because it has the desirable property that the expected value of the criterion conditional on the prediction equals the prediction: $\mathrm{X_r} = \mathrm{E(Y|X)} = \mathrm{E(Y|X_r)}$. Note that, $\mathrm{E(Y \mid X_s)}$ also equals $\mathrm{X_r}$, not $\mathrm{X_s}$. It is not calibrated. In addition to reporting this expected value of Y given $\mathrm{X_r}$ (or $\mathrm{X_s}$), it is often desirable to report a range of Y scores around that predicted score. The upper and lower limits of this range are defined by the standard error of prediction, namely the standard deviation of Y given X, $SD\mathrm{(Y|X)}$, which is inversely related to R, via

$$SD\left(\mathrm{Y|X_s}\right) = SD\,(\mathrm{Y}) * \mathrm{Sqrt}\left(1 - R^2\right),$$

which is placed around the prediction $\mathrm{X_r} = \mathrm{E(Y|X_s)}$.

Table 1 summarizes the conditional expected scores and corresponding score range of the predicted criterion score for different composite predictor scores, $\mathrm{X_s}$. $\mathrm{X_s}$ takes on integer values from 1 to 7, and R takes on four values (1, .5, .25, and .125).

Table 1 Conditional Expected Values and Ranges of the Criterion Scores, Y

Rescaled predicted score (X_s)	$E(Y\mid X_s)$: Range of criterion scores $[E(Y\mid X_s)$: +/− 3*$SD(Y\mid X_s)]$ Y has a mean of 4 and *SD* of 1			
(4, 1) scale	$R = 1$	$R = .5$	$R = .25$	$R = .125$
1	1: [1,1]	2.50: [−.1,5.1]	3.25: [0.35,6.15]	3.625: [0.65,6.60]
2	2: [2,2]	3.00: [0.4,5.6]	3.50: [0.60,6.40]	3.750: [0.77,6.73]
3	3: [3,3]	3.50: [0.9,6.1]	3.75: [0.85,6.65]	3.875: [0.90,6.85]
4	4: [4,4]	4.00: [1.4,6.6]	4.00: [1.10,6.90]	4.000: [1.02,6.98]
5	5: [5,5]	4.50: [1.9,7.1]	4.25: [1.35,7.15]	4.125: [1.15,7.10]
6	6: [6,6]	5.00: [2.4,7.6]	4.50: [1.60,7.40]	4.250: [1.27,7.23]
7	7: [7,7]	5.50: [2.9,8.1]	4.75: [1.85,7.65]	4.375: [1.40,7.35]

Note: Conditional expected values and ranges of the criterion, Y, are on *a scale* with *a mean* of 4 and *a standard deviation* of 1, expressed as a function of rescaled predicted composite score X_s and the correlation (R) between the predicted score and the criterion. There is *only one scale* for Y.

Table 2 Conditional Expected Values and Ranges of the Transformed Criterion Scores, Y_{Rinv}

Rescaled predicted score (X_s)	$E(Y_{Rinv}\mid X_s)$: Range of criterion scores $[E(Y_{Rinv}\mid X_s)$ +/− 3*$SD(Y_{Rinv}\mid X_s)]$ Y_{Rinv} has a mean of 4 and *SD* of $1/R$			
(4, 1) scale	$R = 1$	$R = .5$	$R = .25$	$R = .125$
1	1: [1,1]	1: [−4.2,6.2]	1: [−10.6, 12.6]	1: [−22.8,24.8]
2	2: [2,2]	2: [−3.2,7.2]	2: [−9.6, 13.6]	2: [−21.8,25.8]
3	3: [3,3]	3: [−2.2,8.2]	3: [−8.6, 14.6]	3: [−20.8,26.8]
4	4: [4,4]	4: [−1.2,9.2]	4: [−7.6, 15.6]	4: [−19.8,27.8]
5	5: [5,5]	5: [−.2,10.2]	5: [−6.6, 16.6]	5: [−18.8,28.8]
6	6: [6,6]	6: [0.8,11.2]	6: [−5.6, 17.6]	6: [−17.8,29.8]
7	7: [7,7]	7: [1.8,12.2]	7: [−4.6, 18.6]	7: [−16.8,30.8]

Note: Conditional expected values and ranges of the transformed criterion scores, $Y_{Rinv,}$ are *on scales* with *means* of 4 and *standard deviations* of $1/R$, expressed as a function of rescaled predicted composite score X_s and the correlation (R) between the predicted score and the criterion. There are *four scales* for $Y_{Rinv,}$ one for each level of R.

The first column of Table 1 contains composite predictor score X_s transformed to a scale with a mean of 4 and *SD* of 1, the (4, 1) scale. The next four columns indicate the conditional expected criterion score, X_r, and range of criterion scores in the criterion score (Y) metric around these expected values for selected values of the multiple correlation (R), ranging from a perfect correlation of 1 to a correlation of .125. If the criterion scores were normally distributed, approximately two thirds of criterion scores on the (4, 1) scale scores would fall between 3 and 5, 95% would fall between 2 and 6, and almost all would fall between 1 and 7. Each successive column indicates a halving in correlation. Note that some ranges fall below 1 and above 7.

The column under $R = 1$ lists the expected criterion score and ranges on a 1 to 7 criterion score scale corresponding to each of seven scores on the rescaled composite predictor, which also have a mean of 4 and a standard deviation of 1. Note that when the correlation is perfect, the expected criterion score is identical to the composite predictor value for every composite predictor value. In addition, there is no variation in criterion scores around that expected criterion score; the range of predicted criterion scores is equal to the predicted criterion score. Under this unrealistic situation, an R of 1 produces a predicted criterion value of 1 for a composite predictor value of 1 and a predicted criterion value of 7 for a composite predictor value of 7. This panglossian column is included in Table 1 (and later in Table 2), because an uninformed consumer who does not appreciate the meaning of an imperfect correlation between predicted and criterion scores might assume that the composite predictor value always equals the criterion performance. It never does. It rarely comes close in reality. More realistic expected values (and corresponding ranges) are seen in the last three columns of Table 1 for correlations of .5, .25, and .125, values that might represent rosy, typical, and rather bleak scenarios, respectively, for settings in which supervisory ratings serve as the criterion.

For the average composite predictor value of 4, note that the conditional expected criterion score is 4 and the expected range of rounded predicted criterion scores is [1, 7] for all nonperfect correlations, including $R = .5$ in Table 1. For all but the case of perfect correlation, candidates who have a composite predictor score of 4 could have rounded predicted criterion scores that range from 1 to 7. As the correlation drops to .125, this range of unrounded scores [1.02, 6.98] approaches [1, 7]. Note that the range [1, 7] is the range of scores one would expect given no information about the test taker other than they came from this sample of test takers. In other words, almost all criterion scores would be expected to fall between plus and minus 3 points of the average score of 4. For a composite predictor value of 3, the expected criterion score, rounded to integers, remains 4, and the range of rounded criterion scores is reduced from [1, 7] to [1, 6] for $R = .5$ but remains at [1, 7] for $R = .25$ and $R = .125$.

The rounded reported score range of [1, 7], occurs for all seven composite predictor values when the multiple $R = .125$. It also occurs for composite predictor values of 3, 4, and 5 for $R = .25$. For $R = .5$, it occurs only at a composite predictor value of 4. Correlations as low as .25 and .125 contain a considerable amount of prediction error, and that is reflected in the large ranges of rounded scores associated with given composite predictor scores.

Matching the Predicted Criterion Score Distribution to That of the Criterion

There is another way of interpreting "report predicted scores on the criterion scale." That would involve replacing $\mathrm{X_r}$ with $\mathrm{X_s}$. Note that this substitution of $\mathrm{X_s}$ for $\mathrm{X_r}$ does not affect the correlation (R) between the prediction ($\mathrm{X_r}$ or $\mathrm{X_s}$) with Y. It does, however, alter the relationship between the correlation and the amount of prediction error. When the prediction is $\mathrm{X_r}$,

$$R\left(\mathrm{X_r}, \mathrm{Y}\right) = SD\left(\mathrm{X_r}\right)/SD\left(\mathrm{Y}\right).$$

When we replace $\mathrm{X_r}$ with $\mathrm{X_s}$, the standard deviation of the predictions becomes

$$SD\left(\mathrm{X_s}\right) = SD\left(\mathrm{X_r}\right)/R.$$

To maintain the relationship between the correlation and the ratio of predicted to criterion standard deviations, it is necessary to rescale the criterion Y in the same way that the predictions $\mathrm{X_r}$ were rescaled via $\mathrm{Y_{Rinv}} = (\mathrm{Y} - 4)/R + 4$. This transformation ensures the following relationships:

$$R\left(\mathrm{X_s}, \mathrm{Y_{Rinv}}\right) = SD\left(\mathrm{X_s}\right)/SD\left(\mathrm{Y_{Rinv}}\right),$$

$$= \left(SD\left(\mathrm{X_r}\right)/R\right)/\left(SD\left(\mathrm{Y}\right)/R\right),$$

$$= SD\left(\mathrm{X_r}\right)/SD\left(\mathrm{Y}\right).$$

This transformation is necessary to ensure that the correlation maintains its relationship to a measure of the quality of prediction.

Table 2 summarizes the results of these rescaling operations, which place Y on scales, $\mathrm{Y_{Rinv}}$, with means of 4 and standard deviations of $1/R$. Note the use of the plural is intentional because the transformation is dependent on R and there are four different values of R represented in Table 2, as there were in Table 1. There are four different scales expressed in Table 2, one each for each of the four Rs of 1, .5, .25, and .125.

In Table 1, the original criterion score had a standard deviation of 1, and with almost all scores expected to fall between 1 and 7. The same range of criterion scores occurs in Table 2 only when $R = 1$. A number of entries in Table 2 are difficult to comprehend if the reader presumes that placing the predicted scores on a scale with a mean of 4 and standard deviation of 1 means that the last four columns in Table 2 are a common criterion score scale. They are not. Instead, they are on four very different criterion score scales where the standard deviation of each scale is defined by $1/R$.

For each row in Table 1, the range of criterion scores around the expected score, defined by

$$SD\left(\mathrm{Y}|\mathrm{X_s}\right) = SD\left(\mathrm{Y}\right) * \mathrm{Sqrt}\left(1 - R^2\right),$$

increases as the correlation decreases and the expected score regresses towards 4. In Table 2, the regression effect has been eliminated; $\mathrm{E}(\mathrm{Y_{Rinv}}|\mathrm{X_s})$, is equal to the composite predictor value, $\mathrm{X_s}$: 1 = 1, 2 = 2, 3 = 3. .. regardless of the multiple correlation R.

Another consequence is that the expected range of criterion scores associated with a composite predictor value increases dramatically as the correlation decreases. For each row in Table 2, the range of criterion scores around the expected score, defined by

$$SD\left(\mathrm{Y}_{\mathrm{Rinv}}|\mathrm{X}_{\mathrm{s}}\right) = SD\left(\mathrm{Y}\right) * \mathrm{Sqrt}\left(1 - R^2\right)/R.$$

The effect of reduced correlation on prediction error on the range of values in Table 1, Sqrt($1 - R^2$), has been boosted by the effects of rescaling the criterion by $1/R$.

When R is less than 1, the standard deviation, $1/R$, of rescaled criterion scores ranges from 2 for $R = .5$ to 8 for $R = .125$, with 4 for $R = .25$. In Table 1, we note that knowledge of the composite predictor value when $R = .125$ provides minimal predictive power with respect to the criterion score: the rounded predicted criterion score was always 4 with an expected range of 1 to 7 regardless of the predictor value. In Table 2, however, a 1 is aligned with a 1, and a 7 is aligned with a 7. Note that the range of rounded predicted criterion scores associated with a predictor composite score of 1 has gone from about 6 [1.27, 7.23] in Table 1 to about 48 [−17.8, 29.8] in Table 2 when R is .125; the comparable range for a predictor criterion score of 7 has gone from about 6 [1.40, 7.35] to about 48 [−16.8, 30.8], as noted in the last column.

The cost associated with moving from a predicted criterion score of 4 that ranged from 3.625 for a 1 to 4.375 for a 7 when $R = .125$, to a range of 1 to 7 by matching the predicted criterion score to the criterion score, has been a very large increase in uncertainty of the rescaled predicted criterion score. For $R = .125$, the standard error prediction of .99 on a scale with a standard deviation of 1 is converted to a standard error of 7.94 on a scale with a standard deviation of 8. As a consequence, except for the $R = 1$ column, every range of scores in Table 2 contains at least one number that cannot be obtained on the criterion in its original 1 to 7 metric.

A Recommended Solution

A composite that is based on prediction of a criterion score is a prediction, not a score. It is not a measure of a construct. Instead, it is a composite, whose only reason for existence is prediction of a criterion score.

Continue to assume that a client uses a 1 to 7 scale (with a mean of 4 and an *SD* of 1) such that the ultimate performer is a 7, and a 1 is the opposite. Then assume that the prediction that minimizes the mean squared error of prediction from these eight predictors correlates .33 with the criterion such that the range of these predicted criterion scores goes from 3 to 5. The client could be told that the expected difference in rating between the person with the highest rank and the one with lowest rank on the predicted criterion score is almost two points. The client knows what a 3 and a 5 mean on the 1–7 scale that they routinely use.

If we use a linear transformation of that predicted score such that the new rescaled predicted score standard deviation is equal to the standard deviation of the criterion, the rank order of the original predictions will be unaltered. The 3, however, becomes a 1, and the 5 becomes a 7. Most important, the meaning of the criterion score scale is lost. Unless explicitly instructed to avoid making this mistake, the interpretation that the client attaches to the any numeral may not change even though the meaning of the numeral has changed as a result of the transformation. The client might conclude that the expected value on the criterion scale for a 7 on the predictor scale is a 7, when it fact it is a 5. If presented with conditional ranges of transformed criterion scores, the client might be baffled by scores that take on negative values or any values that fall outside of the criterion score range of 1 to 7. With the help of some complex interpretative material, the client might come to the realization that a prediction of 1 is not a 1 on the criterion scale and that a prediction of 7 is not a 7 on the criterion scale. Using a linear transformation of a predicted score to make the standard deviation of the rescaled predicted scores equal to the standard deviation of the criterion scores obfuscates the meaning of the numbers attached to the prediction of criterion performance.

Whether it is a prediction or a measure of a construct, the scale most readily understood by the client is the scale they use to rank their candidates. The client loses their interpretative hook when the scale they understand is replaced by another one. Flawed inferences about the quality of work they could expect from a candidate selected on the basis of the composite are likely to be made. They will expect too much from candidates with high rankings and too little from those with low rankings. The degree to which their expectations are biased estimates of expected performance of the candidates will be inversely related to the predictive power of the composite.

It is important to present the prediction in a context that is related to criterion performance. The solution is to choose a scale for the criterion score and place the predictions on the scale of the criterion such that the means of predicted and criterion scores are equal and the ratio of the predicted score standard deviation to the criterion score standard deviation equals the multiple correlation associated with the prediction. The simplest course of action is to leave the predicted scores in the metric of the criterion score such that the standard deviation of the predicted criterion score is equal to *R* times the standard deviation of the criterion.

Notes

[1] Although composites such as the ACT composite are used in educational settings, they are not as prevalent there as they are in economics, where indices have been devised to keep track of a variety of complex ideas, such as economic health and economic sustainability. See Stiglitz, Sen, and Fitoussi (2009) for examples.

[2] An exception would be the prediction of a test score on one edition of a test from a score on an earlier edition.

[3] This statement presumes that mean squared error is the loss function that was minimized, as is often the case.

[4] The correlation between the predicted part of UGPA from HSGPA alone and the predicted part of UGPA from SAT (or ACT) score alone is equal to the correlation between HSGPA and SAT/ACT. We are not addressing important questions about the reliability of the test scores or the quality of the criterion or the fact that sample regression equations tend to predict more strongly in the samples they were derived from than in other samples or in the population that the samples represent.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. https://doi.org/10.1007/978-0-387-49771-6

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26–34. https://doi.org/10.1037/0003-066X.48.1.26

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: American Council on Education/Praeger.

Stiglitz, J., Sen, A., & Fitoussi, J.-P. (2009). *Report by the commission on the measurement of economic performance and social progress*. Retrieved from http://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+Commission+report

Suggested citation:

Dorans, N. J. (2018). *Providing a context for interpreting predictions of job performance* (Research Report No. RR-18-38). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12227

Action Editor: Shelby Haberman

Reviewers: Longjuan Liang and Samuel Livingston