Scientific Research Publishing

Psychology, 2025, 16(7), 824-831
https://www.scirp.org/journal/psych
ISSN Online: 2152-7199
ISSN Print: 2152-7180

A Commentary on the Relationship between Pass/Fail Scoring and the Validity of Psychological and Educational Assessments: Using Medical Education as an Example

Clarence D. Kreiter

University of Iowa, Iowa City, USA
Email: clarence-kreiter@uiowa.edu

How to cite this paper: Kreiter, C. D. (2025). A Commentary on the Relationship between Pass/Fail Scoring and the Validity of Psychological and Educational Assessments: Using Medical Education as an Example. *Psychology, 16,* 824-831.
https://doi.org/10.4236/psych.2025.167046

Received: May 12, 2025
Accepted: July 13, 2025
Published: July 16, 2025

Abstract

Background/Introduction: There have been numerous recent initiatives within medical education to convert multi-tiered academic assessments and the grades that summarize them to a dichotomous Pass/Fail (P/F) metric. Because the validity of an assessment depends upon how it is scored, an investigation of the psychometric outcomes of converting a multi-tiered assessment to P/F is needed. **Methods:** The psychometric and statistical validity implications of converting multi-tiered scores and grades to P/F are examined. How P/F scoring impacts test-based learning is also considered. **Results:** For the multi-tiered assessments used in medical education, there will generally be a large decrease in the variance of scores after converting to P/F. Statistical calculations demonstrate that compared to multi-tiered scores, P/F scores exhibit much smaller concurrent and predictive validity correlations and significantly attenuated reliability. Because P/F scores have a limited ability to correlate with other variables and provide very little information (i.e., *low score reliability*) regarding performance, the necessary preconditions for validity are practically absent after converting medical education assessments to P/F. Further, evidence suggests that using assessments with low validity will negatively impact learning. **Discussion/Conclusions:** There does not appear to be a sound psychometric or educational rationale for converting multi-tiered academic assessments to P/F.

Keywords

Pass/Fail Scoring, Reliability, Validity, Educational Assessment, Medical Education, Tests

1. Introduction/Purpose

Across the continuum of assessments used in medical education, there have been numerous initiatives designed to convert multi-tiered academic measures and the grades that summarize them to a dichotomous Pass/Fail (P/F) metric. As the validity of an educational assessment depends upon the scale (i.e., usually a numeric or letter-based system) used to summarize performance, the metrics employed in achievement and aptitude testing will determine the integrity of medical school assessments. Therefore, the score reporting scale needs to be carefully considered. Because the validity of an assessment depends upon how it is scored, medical educators need to understand that changing a multi-tiered score to a dichotomy (i.e., P/F) will influence its measurement characteristics and validity.

When multi-tiered test scores or grades are converted to P/F, the variance is reduced. Since most aspects of construct validity (e.g., generalizability, extrapolation, scoring, implications, prediction, and reliability) depend upon score variance, changing to dichotomous scoring will impact validity. If the information from multi-tiered academic assessments weren't valid, there would be little reason to be concerned about this reduction in variance. However, if academic assessments weren't valid, measurement experts would certainly advise medical educators to eliminate them entirely rather than convert them to P/F. Since academic assessments are generally valid, medical educators must consider how test validity is changed after converting multi-tiered scores to P/F.

This commentary reviews the quantitative evidence that reveals how the psychometric validity of a multi-tiered assessment will change after converting to P/F. It examines how P/F scoring impacts two necessary statistical preconditions for validity. Specifically, it considers how a conversion to P/F influences a test score's ability to correlate with other variables and how the reliability of an assessment is changed after the conversion. Since an assessment's maximum attainable validity is the square root of its reliability (i.e., Max. Validity = $\sqrt{}$reliability), finding a measure to be unreliable conclusively demonstrates low validity. Although finding that a test score is reliable does not necessarily demonstrate it is valid, demonstrating that a test score is unreliable does establish low validity. Therefore, examining reliability as a necessary but not sufficient condition for validity is often a preliminary, and sometimes the final, scientific step in establishing the level of validity exhibited by an assessment score. An additional validity-related concern in educational settings is measuring how an assessment impacts learning.

2. The Statistical/Psychometric Impact of P/F Scoring

The medical education research literature has not adequately assessed the quantitative outcomes of adopting P/F assessments. This is a significant gap in the literature as quantitative statistical methods provide the most scientifically definitive evidence for deciding whether to adopt P/F scoring. In educational measurement, the statistical concepts of reliability and validity represent the scientific definition of measurement quality. Therefore, an estimate of the statistical validity of P/F

scoring is the most important consideration for deciding whether to convert to dichotomous scoring.

2.1. Prediction/Extrapolation

Score variance is reduced when converting multi-tiered ordinal scores (e.g., number correct, percent correct, percentile, decile scores, ordinal letter grades etc.) to a dichotomy (e.g., P/F, 0/1, High/Low). The size of that reduction will depend upon the characteristics of the multi-tiered score distribution and the location of the cut score used to convert a multi-tiered score to a binary metric. In general, the larger the multi-tiered score variance and the smaller the number of students in the failed group (i.e., the further the distance between the P/F cut score and the multi-tiered mean), the larger will be the reduction in score variance upon converting to dichotomous scoring. Given the substantial variance in multi-tiered medical education assessment scores and the small proportion of students who fail (i.e., the relatively large distance between the mean and the cut score), a large reduction in score variance will be observed after converting medical education assessments to P/F.

The reduced variance associated with P/F scoring will also attenuate correlational indices that define score validity. As most are aware, correlation and prediction are inextricably linked to the validity concepts of extrapolation and generalization. When correlation and prediction are reduced, scores will extrapolate and generalize less well. This means that compared to multi-tiered scores, P/F scores will be less valid and less useful for making inferences regarding an examinee's status on the knowledge and skill domains to which a test score generalizes. In other words, P/F scores will be less accurate than multi-tiered scores for establishing the level of competence, characterizing academic achievement, and predicting the future success of students.

In comparing the validity of multi-tiered scores with P/F scores, it is useful to examine their respective correlations with the variables that define their validity. While it may be intuitively obvious that compared to a multi-tiered score, a P/F score will display a reduced correlation with other variables, the magnitude of the difference is best estimated mathematically. Specifically, the impact of dichotomizing an approximately normally distributed assessment score is forecast using **Equation 1** (Cohen, 1983).

$$\rho_{X_D Y} = \rho_{XY}\left(h/\sqrt{pf}\right) \tag{1}$$

Where $\rho_{X_D Y}$ is the projected validity correlation between the dichotomized test score (X_D), and a variable (Y) that is used to quantify test validity, ρ_{XY} is the observed validity correlation between a multi-tiered assessment score (X) and variable (Y), p is the proportion of students who pass, f is the proportion of students who fail, and h is the ordinate of the normal curve at the point of dichotomization. To illustrate, suppose an approximately normally distributed written test score of clinical knowledge is shown to correlate at $r = 0.60$ with examinees' ability to correctly diagnose clinical cases in a performance assessment. If the test is dichoto-

mized to create a P/F exam and 5% of the examinees fail, the correlation with diagnostic performance would be reduced to approximately $r = 0.28$. With this loss of predictive information, the test would be significantly less valid for characterizing diagnostic ability.

2.2. Reliability/Information

How P/F scoring will impact reliability and test information is also an important consideration when deciding whether to convert a multi-tiered score to P/F. The classic representation of reliability is shown in **Equation 2** and provides a useful conceptual tool for representing the proportion of test score variance that conveys meaningful information. **Equation 3** estimates the reliability of a test that is converted from a multi-tiered score to P/F. It provides a perspective on the amount of useful information retained after conversion to P/F. In this equation, ρ_{xxD} equals the reliability of the dichotomized scores, ρ_{xx} equals the reliability of the multi-tiered scores, h is the ordinate of the standard unit normal curve, and p equals the proportion of passing scores. To provide an example, suppose multi-tiered scores display a relative reliability of .70. If a P/F cut score is set at 1.0 SD below the mean (i.e., failing ~16%, $p = 0.841$, $h = 0.242$, $\rho_{xx} = 0.70$), the expected reliability of the P/F score would be $\rho_{xxD} = 0.46$. At 1.5 SD below the mean (i.e., a 7% fail rate, $h = 0.130$), the projected relative reliability would be $\rho_{xxD} = 0.36$.

$$\text{Reliability} = \text{true score var} \Big/ \left(\text{true score var} + \text{error score var}\right) \tag{2}$$

$$\rho_{xxD} = \rho_{xx} * h \Big/ \sqrt{\left[p\left(1-p\right)\right]} \tag{3}$$

2.3. The Outcome of P/F Scoring in Practice

While it is theoretically useful to estimate the reliability as the proportion of observed score variance that is meaningful information and to estimate the predictive validity of a P/F score from multi-tiered score correlations, results from within an actual medical school after converting test scores to P/F provide important practical evidence. A 2016 report by Kreiter and Ferguson (Kreiter & Ferguson, 2016) describes the use of generalizability theory to estimate the reliability of a four-tiered grading system at a large midwestern medical school. They analyzed grades from over 1000 students attending that medical school across 10 consecutive years. They found their four-tiered grading system produced a highly generalizable ($G > 0.80$) (i.e., highly reliable and informative) grade point average (GPA). When those grades were converted to P/F however, the reliability/generalizability was reduced to less than 0.20 ($G < 0.20$). Further, the low reliability of P/F grading substantially attenuated the correlations that defined the validity of those grades (Kreiter & Kreiter, 2007). For example, while GPAs from multi-tiered grades correlate at approximately $r = 0.60$ with licensure scores, a composite GPA based on P/F grades (i.e., the proportion of courses passed on first try) would yield a correlation of just $r = 0.25$ (Kreiter & Platti, 2021). This represents a very large reduction in the validity of those grades. The authors demonstrated that while the

four-tiered grading system used at that school for summarizing performance across assessments did produce a highly reliable and informative GPA, little of that useful information was retained when grades were converted to P/F.

3. The Impact on Learning

Given the importance of acquiring the skills and knowledge needed to practice medicine, a validity investigation of P/F scoring must consider its impact on learning. Testing's ability to facilitate educational achievement is attributed to intrinsic, extrinsic, and selection learning effects (Kreiter et al., 2013). The *intrinsic learning* effect, sometimes referred to as the direct effect of testing, is the degree to which testing or quizzing over studied materials enhances the acquisition and long-term retention of knowledge and skills through the retrieval and rehearsal of information that takes place during testing. The *extrinsic effect*, often referred to as the indirect effect on learning, is the degree to which assessments enhance learning achievements through increased motivation and effort. This takes place when the test scores are used as an accountability mechanism for rewarding the attainment of learning objectives set by the medical college, licensure board, or certification organization. Finally, the *selection effect* is defined as the observed increase in academic achievement by selecting high-aptitude learners with admission tests and other academic achievement measures. While a complete review of the evidence for each effect is beyond the scope of this commentary, existing summaries suggest that each make a substantial positive contribution to learning and that testing represents one of the most powerful interventions employed within medical education (Kreiter et al., 2013; Phelps, 2012; Roediger & Karpicke, 2006; Larsen et al., 2009). Therefore, the relevant validity question is how converting multi-tiered scores to P/F impacts these types of test-based learning effects. Although there is little experimental evidence (e.g., *randomized controlled trials)* in medical education that applies to this question, other forms of evidence offer insight. That evidence is presented in the following three sections.

3.1. The Intrinsic (Direct) Effect

Correct answer formative feedback and the mental practice of producing a correct answer during testing has been shown to produce substantial gains in learning compared to studying alone (Roediger & Karpicke, 2006; Larsen et al., 2009). Assessments used purely as intrinsic learning tools are rare in medical education. This is because a test designed solely for intrinsic learning is, by definition, a no-stakes assessment with no external consequences. For such tests, learning would be maximized by informing the examinee regarding the quality of each response. Clearly this requires scoring at the item level. However, summary P/F scoring across items in this context might be a useful indicator of mastery and setting a logical stopping point to terminate an intrinsic test. In this application, an additive cut-score across dichotomous items to indicate mastery would appear to be a valid use of P/F scoring. This application would be primarily useful as a programming

technique in computerized automated test-based learning as a method for defining a termination point for self-testing. However, more traditional and much more frequently employed stakes-based assessments in medical education also contribute to intrinsic learning through practice and formative feedback. If the distance a particular performance falls from the P/F cut score is useful for intrinsic learning, logic suggests that its elimination would diminish the intrinsic effect by removing information regarding the quality of a test performance. In other words, if it is useful feedback for an examinee to know how their performance ranks relative to their peers or how far their score exceeds or falls below the passing mark, then P/F scoring will be detrimental to the formative process of informing the learner whether they have sufficiently mastered the content.

3.2. The Extrinsic (Indirect) Effect

The extrinsic or indirect learning effect is quantified by comparing the level of student achievement with and without the accountability mechanisms enabled by assessments. Assessments that promote this effect are course-based tests, course grades, licensure examinations, and certification examinations. The indirect effect is enabled by test scores that convey the relative or absolute amount of learning a student has achieved. Assessments are often used to increase student efforts by allocating rewards or penalties based upon assessment performance. Studies that report a reduction in self-reported student stress with P/F grading could be interpreted as consistent with the hypothesis that P/F grading lowers extrinsic motivation and the amount of study time a student is willing to dedicate to their course work. While research has examined the correlation between the use of examinations and the subjective experience of stress, stress and effort are confounded, if not identical, constructs in such studies. The true quantitative impact of P/F scoring on the extrinsic effect has not been adequately examined by a randomized controlled trial design that compares medical education programs with and without P/F scoring. Although several within institution pre-post studies have shown little or no effect on average licensure scores before and after converting to P/F grading, multi-tiered test scores (e.g., norm-based or percent correct scores) continued to be used for rewarding high levels of performance in those studies (Spring et al., 2011).

3.3. The Selection Effect

The score on an admission test is one of the first academic measures obtained from a medical student. Admission measures were also among the first to be considered for conversion to P/F reporting. An early call to convert the *Medical College Admission Test* (MCAT) to P/F came from a previous President of the AAMC, Jordan Cohen, who advocated removing MCAT scores from an applicant's file and providing admission committees with only information regarding whether the score was above or below a passing threshold (Cohen, 2002). This recommendation was based upon the belief that P/F scoring of the MCAT would improve the

so-called 'non-cognitive' attributes of the admitted applicants and enhance diversity. Unfortunately, this practice can be shown to yield large decreases in average achievement as reflected by predicted licensure scores for both the over and under-represent groups (Kreiter, 2007). In addition, no evidence has ever been reported showing a negative correlation between the favored non-cognitive attributes and cognitive aptitude. This implies there is no reason to assume there would be an improvement in non-cognitive attributes with P/F scoring of the MCAT. The most compelling reason for retaining multi-tiered admission test scores however is reflected by meta-analytic analyses and statistical modeling of the selection-effect. Studies examining the impact of selection measures clearly show that selection using multi-tiered aptitude and achievement measures provide large gains in learning (Effect Size > 1.0) and that converting multi-tiered admission metrics to P/F would attenuate learning outcomes to a significant degree (Kreiter et al., 2013; Kreiter, 2007).

4. Summary and Conclusions

Psychometric and statistical evidence demonstrate that correlation and reliability are significantly reduced by converting multi-tiered scores to P/F. This implies that P/F scoring will negatively impact test validity. Statistical metrics measuring information, extrapolation, prediction, and generalizability will be substantially compromised with P/F scoring of most medical education assessments. Except for tests designed for purely intrinsic learning, the evidence suggests that the necessary conditions for validity are almost completely absent for assessments utilizing P/F scoring. Further, administering tests with very low validity will have a negative impact on learning. From a scientific perspective, it is illogical to expect that medical education would improve simply by removing information regarding an examinee's performance on an assessment. Rather, statistical evidence shows that multi-tiered scores are more reliable and valid and better able to promote learning. Although medical students have expressed a preference for P/F scoring in some circumstances, it is unknown whether they understand that P/F scoring reduces validity, reliability, and test-based learning. Given this, it would be useful to assess student preferences after informing them of the measurement and learning impacts of P/F.

Recent decisions to remove test score information based upon the belief that it will improve medical education appear misguided. While a reduction in stress appears to be a commonly observed outcome as reflected in subjective evaluations (Rohe et al., 2006), effective education relies on accountability. While accountability could be viewed as introducing stress, removing accountability measures would certainly reduce effort as well as stress. Removing sound measures to achieve demographic representation is also misguided as it removes measurement precision for both the over- and under-represented groups. Alternately, employing affirmative action with multi-tiered scores maximizes within group predicted performance. Differential standards achieve a higher level of public protection and are more trans-

parent. In addition, the assertion that multi-tiered licensure scores mislead medical educators and students appears untenable. Such claims fail to account for the important role of reliability and validity and underestimate the intellectual capabilities of test score users. Users of medical education test results are quite capable of understanding and using the test statistics along with the interpretative guidance that accompanies multi-tiered score reports. Compared to the option of trusting test users with valid and reliable multi-tiered test score information, it is difficult to support the ethics of withholding that information and providing only a P/F score.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Cohen, J. (1983). The Cost of Dichotomization. *Applied Psychological Measurement, 7,* 249-253. https://doi.org/10.1177/014662168300700301

Cohen, J. J. (2002). Our Compact with Tomorrow's Doctors. *Academic Medicine, 77,* 475-480. https://doi.org/10.1097/00001888-200206000-00002

Kreiter, C. D. (2007). A Commentary on the Use of Cut-Scores to Increase the Emphasis of Non-Cognitive Variables in Medical School Admissions. *Advances in Health Sciences Education, 12,* 315-319. https://doi.org/10.1007/s10459-006-9003-9

Kreiter, C. D., & Ferguson, K. J. (2016). An Investigation of the Generalizability of Medical School Grades. *Teaching and Learning in Medicine, 28,* 279-285. https://doi.org/10.1080/10401334.2016.1154859

Kreiter, C. D., & Kreiter, Y. (2007). A Validity Generalization Perspective on the Ability of Undergraduate GPA and the Medical College Admission Test to Predict Important Outcomes. *Teaching and Learning in Medicine, 19,* 95-100. https://doi.org/10.1080/10401330701332094

Kreiter, C. D., & Platti, N. L. (2021). The Medical School Grade Validity Research Project: Grade Reliability. *Health Education and Public Health, 4,* 387-392.

Kreiter, C. D., Green, J., Lenoch, S., & Saiki, T. (2013). The Overall Impact of Testing on Medical Student Learning: Quantitative Estimation of Consequential Validity. *Advances in Health Sciences Education, 18,* 835-844. https://doi.org/10.1007/s10459-012-9395-7

Larsen, D. P., Butler, A. C., & Roediger III, H. L. (2009). Repeated Testing Improves Long-Term Retention Relative to Repeated Study: A Randomised Controlled Trial. *Medical Education, 43,* 1174-1181. https://doi.org/10.1111/j.1365-2923.2009.03518.x

Phelps, R. P. (2012). The Effect of Testing on Student Achievement, 1910-2010. *International Journal of Testing, 12,* 21-43. https://doi.org/10.1080/15305058.2011.602920

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science, 17,* 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rohe, D. E., Barrier, P. A., Clark, M. M., Cook, D. A., Vickers, K. S., & Decker, P. A. (2006). The Benefits of Pass-Fail Grading on Stress, Mood, and Group Cohesion in Medical Students. *Mayo Clinic Proceedings, 81,* 1443-1448. https://doi.org/10.4065/81.11.1443

Spring, L., Robillard, D., Gehlbach, L., & Moore Simas, T. A. (2011). Impact of Pass/Fail Grading on Medical Students' Well-Being and Academic Outcomes. *Medical Education, 45,* 867-877. https://doi.org/10.1111/j.1365-2923.2011.03989.x