Health Education and Public Health
2021; 4(1): 387 – 392 doi: 10.31488 /HEPH.160

Research Article

The Medical School Grade Validity Research Project: Grade Reliability

Clarence D. Kreiter, PhD[1], Nicole L. Platti, BS, M2

University of Iowa Carver College of Medicine, USA

[*]Corresponding author: Clarence D. Kreiter, University of Iowa Carver College of Medicine, 1204 Medical Education Building, Io wa City, Iowa 52242, USA.

Received: January 27, 2021 ; Accepted: February 03, 2021; Published: February 17, 2021

Abstract

Introduction / Purpose: Grading in medical school is currently unstandardized. A consensus regarding the best approach for summarizing performance is difficult to achieve without an understanding of the validity evidence for grades. This paper identifies six validity questions that need to be addressed for deciding how best to summarize student performance in medical school. This research goes on to address the first of those question regarding grades reliability. Method: A broad literature search using Prisma guidelines was conducted. The literature was reviewed and the studies providing quantitative evidence of medical school grade reliability were retrieved. The estimates were entered into a meta-analysis for both didactic and clinical courses. Results: One direct estimate and eight indirect estimates from over 5000 medical students and eight medical schools yielded lower bound estimates of reliability of .70 for didactic one-year GPA and .44 for clinical courses. When the indirect estimates were adjusted with a realistic true score correlation, the reliability estimates were .82 and .66 for one-year didactic and clinical GPA respectively. Discussion / Conclusion: The existing literature suggests grades are sufficiently reliable measures of student performance to support medium to high stakes decisions.

Keywords: The medical school grade validity research project, grade reliability, grades, medical education, validity, reliability.

Introduction

Several years ago, while staffing a research poster characterizing the reliability of medical school grades [1], a newly minted resident approached to consider our findings. When he finished reading, the recent medical school graduate asked whether I (CK) considered grades useful. I acknowledged that while the existing literature did not offer a systematic consideration of medical school grade validity, the research summarized on the poster did suggest that grade point average (GPA) provides a precise (reliable) indicator of student achievement at our medical school. The resident expressed surprise upon learning that a scientific consideration of grade validity did not exist and wondered aloud how his medical school had come to adopt their grading plan. He went on to describe his experience as a medical student and how it shaped his perspective on grading.

The resident related that he had attended a medical school that had recently eliminated multi-tiered grades and adopted Pass/Fail (P/F) reporting. He further described how his dedication to the study of medicine had led to long hours in the library and deferred social opportunities - but did allow him to receive top scores on his exams and assignments. Unfortunately, as his education progressed, he became increasingly disillusioned with each course 'Pass' he received to summarize those accomplishments. Apparently, his roommate, who spent very little time studying, repeatedly missed class, and scored poorly on his tests and assignments, had received the same 'Pass' marks. Listening to this resident's personal experience, it was easy to understand his disappointment upon realizing that his roommate's approach to medical school had produced a transcript identical to his own. Perhaps more importantly, he confided that after coming to understand that his school's reporting would not distinguish between widely differing performances, he did decide to reduce the time he allocated to the school's curriculum.

Although, at the time, I was unable to impart much insight or solace to this resident, our interaction did provide insight into a recent graduate's perspective on grading. It also prompted a more careful consideration of his questions regarding the usefulness of grades and how medical schools might decide on the best approach for summarizing achievement. Upon considering the current status of the literature, we realized that an answer to these questions was possible only with a careful consideration of the validity evidence.

To determine what evidence might be useful in establishing the validity of medical school grades, it is important to consider how our current reporting practices evolved. The literature shows that while there is general agreement on the methods for, and importance of, valid and reliable course assessment, medical educators have not developed a consensus on how best to report and summarize course performance or whether the reliability and validity of summary measures matters. Although traditional multi-tiered grading approaches (i.e. A, B, C, D, F and various numbering systems) have a long history of use in medical education, over the last decades educators have questioned the metric and role of traditional grades and many medical schools have abandoned their use [2,3]. In editorials and informal reviews from as early as the 1960s, medical educators have voiced skepticism regarding the validity, reliability and the impact of grades on student learning and psychological health [4,5]. Decisions to transition away from multi-tiered grades and adopt P/F-style reporting typically cite validity concerns related to student wellness, collaboration over competition, and the need to enhance intrinsic learning goals. Medical students have also shown support for P/F reporting with petitions to eliminate grades at a number of programs across the country. The movement away from traditional grading is reflected in a 2012 survey of 119 of 123 LCME-accredited AAMC-member medical schools that found just 22 (17%) United States (U.S.) medical schools continued to use traditional grades for clerkships [6]. The other 97 (83%) programs employed eight different metrics with a range of 2 to 11 tiers (scale points). Even for schools sharing a common number of tiers, the definitions for those tiers were often quite different. A perspective on pre-clerkship grading trends is provided by the LCME Medical School Questionnaires administer across the 2015 - 2019 academic years [7]. It revealed that the number of U.S. medical schools utilizing P/F course reporting increased from 87 to 116 over that time period and that the other 60 schools utilized five different tiered reporting systems to summarize course performance in 2019.

Problems arising with our current course reporting systems are documented in a report from the University of California by Osborn et al. [8]. The authors recount their difficulties in fairly and objectively interpreting medical student performance evaluations that display a wide array of unstandardized formats. These sorts of interpretative problems are likely to become even more problematic in 2022 when the USMLE Step 1 transitions to P/F reporting and further reduces the objective information regarding medical student performance. Addressing issues related to course reporting is not possible without a systematic consideration of medical school grade validity. Without an understanding of the validity evidence, claims regarding grade accuracy, grade precision and the impact on learning and student well-being are impossible to confirm and a consensus regarding a standardized approach difficult to achieve.

Given the current and past expressed dissatisfaction with grading, it is somewhat surprising that a comprehensive study of medical school multi-tiered grade validity has never been reported. Without a systematic consideration of that evidence, it is not possible to judge whether concerns over traditional grades are well- founded, or to determine the comparative success of the new methods that have replaced them. For example, to answer the resident’s question regarding the wisdom of switching to a P/F metric, it is necessary to first understand the measurement characteristics of the multi-tiered grades they replaced. To address this, we initiated the Grade Validity Research Project. This is the first in a series of six research studies designed to systematically evaluate the validity of grades awarded at U.S. medical schools. Ultimately, the validity of measures used in medical education depends upon whether they promote the educational mission and whether they facilitate defensible inferences for making educational and professional practice decisions. Kane’s four validity inferences (generalization, scoring, extrapolation, and implication) are well suited for structuring such an inquiry [9]. They are used to generate the six research questions displayed in Table 1 along with the specific inference and the type of validity evidence. This report addresses research Question 1.

Table 1. Grade validity research questions.

Validity Question	**Type of Evidence**	**Relevant Inference**
Question 1: **Are medical school grades reliable?**	**Reliability and Internal Structure**	**Generalization**
Question 2: Do grades accurately reflect academic achievement in medical school?	Content	Extrapolation
Question 3: What scale or metric is best for summarizing medical student performance?	Response process and Internal Structure	Scoring
Question 4: Do medical school grades provide unique information beyond that provided by medical licensure examinations?	Relation to Other Variables	Scoring
Question 5: How are grades related to educational and professional outcomes?	Relation to Other Variables	Extrapolation
Question 6: How do grades impact learning in medical school?	Consequence	Implications

Method

For each of the six questions, meta-analytic techniques will be used to summarize the existing literature. In addressing the first question, we initially conducted a broad systematic literature search using Prisma guidelines [10] that included any study investigating health science education or medical school grades. We retrieved each title and abstract from that broad search for review and assessed whether the article might provide quantitative evidence for addressing the question of medical school grade reliability. Since the first stage of the search was designed to include the literature addressing validity more broadly, the inclusion criteria placed no limit on the type of information reported, which specific health professions were included (i.e. medical, nursing, dental, PT, OT, etc.), the publication date or the study location.

As displayed in Figure 1, the initial search for studies reflecting on grades yielded 6528 total articles: PubMed = 1450, Embase = 2034, CINAHL = 1132, ERIC = 1834, and Cochrane CENTRAL = 78. After eliminating duplicates, 4475 articles remained. The titles and abstracts were screened independently by two of the authors (NP and CK) and another 4409 were excluded as they did not address grade reliability. The remaining 66 full-text articles were then closely assessed for eligibility. At this stage we excluded research reports where a reliability coefficient or other quantitative evidence for deriving reliability was not provided. After reviewing the relevant results, it was determined that for addressing the question of medical school grade reliability, the literature related to medical student grading (the focus of our research) was sufficient to obtain a reasonably accurate and generalizable estimate of medical school grade reliability. This allowed us to exclude the less relevant research literature from the other health professions in our summary analysis. For estimating medical school grade reliability, quantitative data (e.g. alpha coefficients, generalizability coefficients, and correlation coefficients) were extracted for use in the meta-analytic summary.

Results

Table 2 displays the selected research. One article directly estimated medical school grade reliability. That study, conducted at a large midwestern U.S. state medical school, was based on a generalizability (G) analysis of grades from 1001 students over 10 academic years (2002-2012) [1]. That research found that a G coefficients G = 0.88 for a GPA summarizing one year of study for didactic grades and a G = 0.77 for a one-year clinical GPA. In the remaining eight studies a lower limit estimate of grade reliability could be derived from its relationship with the USMLE Step 1 scores, a variable of known reliability. The reliabilities of GPA were calculated using the attenuation for reliability equation:

equation: $\rho_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}}$, where:

ρ_{xy} = true-score correlation .

r_{xy} = observed correlation

r_{xx} = reliability of GPA, and

r_{yy} = reliability of USMLE Step 1 (0.85)

We conservatively set the true-score correlation to ρ_{xy} = 1.0 to solve for r_{xx}.

Data for GPA reliabilities are summarized in Table 2.

For the studies where a lower bound estimate of medical school grade reliability could be derived, the samples were drawn from 7 medical schools using the multi-tiered grades received by 5623 students across the U.S. For the most conservative (ρ_xy = 1.0) average derived one-year didactic GPA reliability an r_{xx}= 0.70 [n = 5,623] was obtained and an r_{xx}= 0.44 [n = 3,779] for the clinical one-year GPA. When the true score correlation for didactic was set at .90 and .80 for clinical, the reliability estimates were .82 and .66 respectively.

Table 2. Reliability Estimates for One-Year Didactic and Clinical GPA Assuming True Score Correlation (TSC) = 1.0 – {assuming TSC = .90 Didactic & TSC = .80 Clinical}

Study	Study Characteristics Derived vs. Direct [n subjects] – Scale	Rel. GPA One Year Didactic (D) / Clinical (C)
[1]	Direct [n = 1101] – 4 pts	0.88 (D) 0.77 (C)
Paolo, Bonaminio et al. (2004) (11)	Derived [n = 686] - 5 pts	0.64 (D)
Gandy, Herial et al. (2008) (12)	Derived [n = 711] – Percent	0.87 (D)
Andriole, Jeffe et al. (2005) (13)	Derived [n = 237] – 5 pts	0.24 (C)
Sesate, Milem et al. (2017) (14)	Derived [n = 96] – Percent	0.74 (D)
Denton, Durning et al. (2010) (15)	Derived [n = 588] – 5 pts	0.61 (D)
Zahn, Saguil et al. (2012) (16)	Derived [n = 484] – 5 pts	0.66 (D) 0.36 (C)
Dong, Saguil et al. (2012) (17)	Derived [n= 802] – 5 pts	0.58 (D) 0.58 (C)
During, Dong et al. (2015) (18)	Derived [n = 1155] – 5 pts	0.60 (D) 0.26 (C)
Totals and Averages	**Total N = 5,623**	**0.70 (D) {.82} (D)**
	Total N = 3,779	**0.44 (C) {.66} (C)**

Discussion

The results from the meta-analysis across eight medical schools and from over 5000 medical students provided reasonably consistent findings. For courses employing didactic instruction, the average reliability of GPA across one year of study was estimated to be r_{xx} =0.70 - 0.82. For clinical courses / clerkships the reliability for a GPA summarizing one year of study was r_{xx} = 0.44 – 0.66. In interpreting these results, it is important to keep in mind that for the eight studies where the reliability for GPA was derived, the most conservative reported results were certainly an underestimate. This is because we implemented maximum caution in our estimates by setting the true score correlation between Step 1 and grades to p_{xy}= 1.0. It seems much more likely that a true score correlation of 0.90 depicts the actual state of the relationship. This underestimation is likely to have been especially pronounced for the clinical courses as Step 1 primarily assesses non-performance knowledge-related learning that takes place during didactic instruction. With an assumed true score correlation of p_{xy} = 0.90 for the didactic and p_{xy}= .80 for the clinical, the average one-year didactic grade reliabilities exceed r_{xy}= 0.80 and the average one-year clinical GPA reliabilities are greater than r_{xx} = 0.65. Given this, it seems safe to conclude that these measures are among the most reliable indicators we obtain regarding student performance. In addition, since didactic and clinical GPAs are often calculated and reported as composites across multiple years and sometimes as a composite across clinical and didactic years, the observed reliability for two or more years GPA would be considerably higher than the one-year reliabilities reported here.

Interestingly, one paper in our search reported on a comparison between a multi-tiered grading approach and P/F-type reporting system .Since a reliability statistic can be interpreted as conveying the proportion of score variance that reflects meaningful information about a student (much like a signal-to-noise statistic), that study compared the information contained in a 4-tier GPA measure with two-tiered P/Fre porting. The researchers concluded that almost all useful information was lost with P/F reporting. Of course, this is not surprising given that very few students (< 2%) received a failing grade. If the cut score defining failure was set higher, the information (reliability) would certainly increase somewhat. But, of course, failing more students to attain better reliability cannot be recommended

Finally, although the literature contains only one direct estimate that allowed the most precise approach to estimating medical school grade reliability, the addition of the derived estimates did allow us to confidently conclude that GPA can, and usually does, reach the levels of reliability that are needed for making medium to high- stakes decisions.

Compliance with Ethical Standards
Funding: NA
Conflict of Interest: NA
Ethical Approval: NA Informed
consent: NA

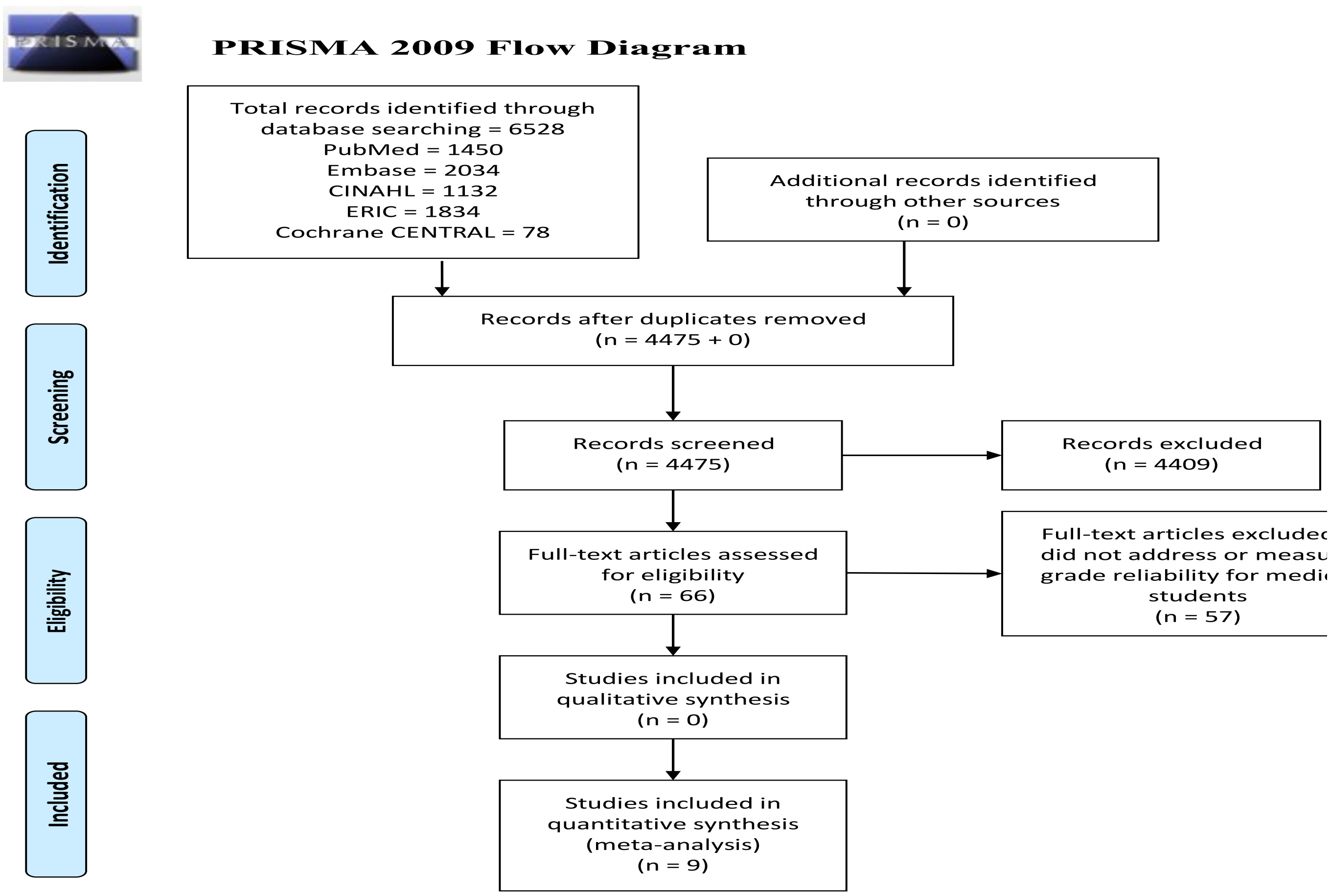
PRISMA
PRISMA 2009 Flow Diagram
Identification
Screening
Eligibility
Included
Total records identified through database searching = 6528
PubMed = 1450
Embase = 2034
CINAHL = 1132
ERIC = 1834
Cochrane CENTRAL = 78
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = 4475 + 0)
Records screened (n = 4475)
Records excluded (n = 4409)
Full-text articles assessed for eligibility (n = 66)
Full-text articles excludec did not address or measu grade reliability for medi students (n = 57)
Studies included in qualitative synthesis (n = 0)
Studies included in quantitative synthesis (meta-analysis) (n = 9)

Figure 1. PRISMA 2009 flow diagram.
From: Moher D, Liberti A, Tetzlaff J, Atman DG. The PRISMA Group (2009). Preferred Reporting for Systematic Reviews and Meta Analyses: The PRISMA Statement. PLos Med 6(6): e1000097.doi: 10.137/journal. pmed1000097.

References

1. Kreiter CD, Ferguson KJ. An Investigation of the Generalizability of Medical School Grades. Teaching and Learning in Medicine, 2016;28(3):279-285.
2. Hughes RL, Golman ME, Patterson R. The grading system as a factor in the selection of residents. J Med Edu.1983;58:479-481.
3. Magarian GJ, Mazur DJ. A national survey of grading sytems used in medical clerkships. Academic Med. 1990;65:636-639.
4. Wingard JR, Williamson JW. Grades as predictors of physician’s career performance: An evaluative literature review. J Med Edu.1973;48:311-320.
5. Stimmel B. The use of pass/fail grades to assess academic achievement and house staff performance. J Med Educ.1975;50:657-651.
6. Alexander EK, Osman NY, Walling JL, et al. Variation and imprecision of clerkship grading in U.S. medical schools. Academic Med. 2012;87:1070-6.
7. Association of American Medical Colleges. Medical School Graduation Questionnaire: 2015-2019 All Schools Summary Report.
8. Osborn MG, Mattson J, Yanuck J, et al. Ranking practice variability in the medical school performance evaluation: so bad, it’s “good”. Academic Med. 2016;91(11):1540-1545.
9. Kane, M.T. Validating the interpretations and uses of test scores. J Edu Measurement 2013 50:1-73.
10. Moher D, Liberati A, Tetzlaff J, et al. The PRISMA Group (2009). Preferred Reporting Items for systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoSMed. 6(60: e1000097.
11. Paolo AM, Bonaminio GA, Durham D, et al. Comparison and cross-validation of simple and multiple logistic regression models to predict USMLE step 1 performance. Teaching Learning Med. 2004;16(1): 69-73.
12. 12. Gandy RA, Herial NA, Khuder SA, et al. Use of Curricular and Extracurricular Assessments to Predict Performance on the United States Medical Licensing Examination (USMLE) Step 1: A Multi-Year Study. Learning Assistance Rev. 2008;13(2): 27-35.
13. Andriole DA, Jeffe DB, Hageman HL, et al. What predicts USMLE Step 3 performance? Academic Med. 2005;80(10 Suppl): S21-24.
14. Sesate DB, Milem JF, McIntosh KL, et al. Coupling Admissions and Curricular Data to Predict Medical Student Outcomes. Res Higher Edu.2017;58(3): 295-312.
15. Denton GD, During SJ, Wimmer AP, et al. Is a faculty developed pretest equivalent to pre-third year GPA or USMLE step 1 as a predictor of third-year internal medicine clerkship outcomes?" Teaching and Learning in Med. 2010;16(4): 329-332.
16. Zahn CM, Saguil A, Artino AR, et al. Correlation of National Board of Medical Examiners scores with United States Medical Licensing Examination Step 1 And Step 2 scores. Academic Med. 2012;87(10): 1348-1354.
17. Dong T, Saquil A, Artino AR, et al. Relationship between OSCE scores and other typical medical school performance indicators: a 5-year cohort study. Military Med.2012;177(9 Suppl): 44-46.
18. Durning SJ, Dong T, Hemmer PA, et al. Are commonly used premedical school or medical school measures associated with board certification? Military Med. 2015;180:18-23.

To cite this article: Kreiter C, Platti NL. The Medical School Grade Validity Research Project: Grade Reliability Health Education and Public Health. 2021; 4:1.