CollegeBoard
connect to college success™

Research Notes
Office of Research and Development RN-20, October 2004

Consistency and Reliability in the Individualized Review of College Applicants

Emily J. Shaw and Glenn B. Milewski

Introduction

Reliability refers to the consistency and stability of a measure from one use to the next (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). While all measures contain some amount of measurement error, an unreliable measure contains too much measurement error. For example, if you got on a scale and it showed that you weighed 150 pounds one minute and 138 pounds the next minute, and repeating this process resulted in readings of very different weights each time, then the scale would not be reliable and would contain a great deal of measurement error (Vogt, 1999). If, however, you repeatedly weighed yourself on the same scale and each time it read 145 pounds, your scale would be considered reliable, although it might not necessarily be accurate.

For the purposes of individualized review in the college admissions process, reliability becomes a major concern when a number of different readers evaluate and make important recommendations or actual decisions based on somewhat subjective application materials (Rigol, 2003). Individualized reviews of college applicants focus not only on academics (grades and test scores) but also on the applicant's talents, experiences, and potential (as measured by personal statements, letters of recommendation, high school activities, etc.). In the case of individualized reviews, the concern is not as much with the reliability of the applicant's essay, SAT® score, or high school grades over time, but rather it is with the reliability of the application ratings. In other words, the focus is on the consistency of ratings of admissions materials between two or more readers or by different readers in settings where only one reader rates an application. When a file is reviewed by only one reader, and different readers are responsible for a certain number of files, the concern arises that some readers may be more lenient or stringent than others when making judgments about the applicants' qualifications. If reader ratings or decisions are unreliable, it is likely that when an application is reviewed by another reader, the new reader's rating and decision will be different from a previous rating or decision.

Interrater reliability refers to the agreement between readers, or the extent to which readers judge or rate a college application, performance, or production in the same way (Vogt, 1999). There are several aspects of interrater reliability. The first is the *composite reliability* of judges or readers; it can be evaluated by correlating ratings made by different readers on the same group of applicants. The second is *reader consistency*; it can be evaluated by calculating the percent agreement between different ratings on the same group of applicants. A third aspect of interrater reliability is *interrater severity,* which captures the degree of leniency or stringency of different readers by comparing average ratings between them. Each aspect of interrater reliability is important for a college or university to evaluate when reviewing college applications.

Interrater Reliability in Individualized Reviews

Composite reliability is calculated by correlating scores assigned by two or more independent readers on an appropriately selected sample of applications. The correlations between scores show whether the readers tend to give high or low ratings in a consistent manner (Rubin and Babbie, 1993). A statistical adjustment (the Spearman-

Brown formula) is applied when estimating composite reliability to account for the number of readers:

$$r_{11} = \frac{n\bar{r}}{1 + (n-1)\bar{r}}$$

where r_{11} is the composite reliability, n is the number of readers, and $\bar{r}$ is the average correlation among readers. The formula provides a value that ranges from –1 to 1, where 1 reflects perfect reliability.

Reader consistency can be examined by calculating the proportion of times that applicants' admissions materials receive exactly the same scores from a pair of readers and/or the proportion of scores that fall within ±1 point of each other (Linn and Gronlund, 2000). For example, imagine that two readers are given 100 applications to rate on a three-category checklist. The first category is for applications that are "not qualified," the second category is for applications that are "questionably qualified," and the third category is for applications that are "definitely qualified." If the two readers checked the same category for 90 of those applications, then the percent of agreement between readers would be 90 percent. It is important to note that percent agreement can be a misleading index because it fails to differentiate between accuracy and variability (see Rosenthal and Rosnow, 1991, pp. 54–55, for a more complete review).

Interrater severities (Weigle, 1998) provide another valuable source of information, particularly for instances when an application is examined and rated by only one reader and the reader will not be reviewing every application. Severities can be represented by the average scores assigned by different readers for either the same applications or for all of the applications they have assessed over time (Linn and Gronlund, 2000). It is important to compare the average scores of each reader in order to check whether there is a strong tendency for one reader to be consistently more or less lenient than another (Weigle, 1998). Such a situation might result in some applications receiving higher scores because a reader is more generous and some applications receiving lower scores because a reader has higher expectations. Estimates of reader severity can provide insight into whether there is a need for further work "calibrating" the readers.

It is important to note that some variation in ratings of college applications is expected since no reader is completely consistent. However, this variation should not be unduly influenced by measurement error. Sources of measurement error can be thought of as either internal or external to the reader (Rosenthal and Rosnow, 1991). Internal sources of measurement error may include the reader's level of motivation, interest, attention span, or health, all of which can affect the reliability of the reader's application ratings. Fatigue is one common source of internal error experienced by readers because they typically review hundreds of applications in a short period of time. Measurement error that is external to the reader may include variation in the amount of training readers receive or specificity of the rubrics they use, for example. Effects of reader subjectivity and variation in reader standards both play a role in interrater reliability.

Evaluating interrater consistency is very important to any application that must be judgmentally scored. In an individualized review process, not only are the applicants receiving judgmental scores or ratings, but these scores become part of important decisions. It should be noted, however, that interrater reliability does not take into account how consistently an applicant performed on the essay, the academic transcript, the activities or community involvement, or the interview. Consistency across different tasks most resembles internal consistency reliability, which is used to determine the reliability of a set of test items.

Encouraging and Improving Consistency and Reliability

One helpful way to encourage reliability between readers is to have the readers meet somewhat regularly to discuss their ratings of several of the same applicants and their reasoning behind the scores they assigned. If disagreements arise, the readers can discuss them and arrive at rules or guidelines for assigning particular scores on different parts of the application (essay, letters of recommendation, academic transcript).

Reader training is another good method to improve reliability and reduce measurement error, especially for assessment procedures that require subjective judgments to be made on constructed responses. This training would likely require the participation of admissions staff, professors, alumni, or anyone who is involved in the reading process. The training might focus on informational and/or practice sessions aimed at identifying and agreeing upon the constructs that are being assessed in the individualized review, as well as how these constructs can/should be most appropriately measured. Training has been found to increase reader self-consistency, though it is not necessarily the most effective way of eliminating differences in the amount of leniency or stringency in scoring; it does, however, seem to bring the extreme scorers within a more tolerable range of severity (Weigle, 1998). Major differences in severity that arise in the individualized review process will likely require significant dialogue between all readers involved as this may be the result

of differing definitions of the construct that the application is intended to measure.

Rubrics are another way to improve interrater reliability. Rubrics facilitate reader agreement by explicitly outlining the standards or achievements that correspond to different ratings. Rubrics assist in assigning levels of achievement to student-produced material; they usually consist of ordered categories coupled with descriptions of criteria that match those categories (Schafer, Swanson, Bené, and Newberry, 2001). Scoring criteria in rubrics should reflect the content and processes judged by the admissions committee to be important. Creating well-defined, detailed rubrics requires the college or university to make clear value judgments and determine the most important and critical aspects of performance, achievement, and potential that the school is looking for in an applicant (Parke, 2001). For example, a rubric used to review an individual's extracurricular activities, service, and leadership may include categories such as awards and honors received, community service, and leadership positions to be rated on a 1–10 scale. This can help "standardize" the process, enhancing consistency as well as explicitly defining what is important to the institution. Therefore, the decision of what to include in a rubric for the individualized review process should be deliberate and well thought out.

In an effort to achieve high interrater reliability on the SAT essay, the College Board chose to implement rigorous reader training, detailed scoring rubrics, and the use of reader calibration. Although rating SAT essays is not the same as rating college applications, it is a situation where student-produced materials are evaluated by many different readers. Each SAT essay (there are approximately 2.5 million each year) will be read by two readers. If the two readers' scores differ by more than one point, a third reader will score the essay. The third reader will be an experienced reader with special training in holistic scoring; this score will function as the test-taker's final rating. The College Board expects that more than 92 percent of all scored essays will receive ratings within ±1 point of each other on the 6-point SAT essay scale (College Board, 2004).

Interviews

In the context of college admissions, interviews are sometimes conducted to gather information to guide admissions decisions. Muchinsky (1987) explains that interviews can range from highly structured, where the interviewer asks each applicant a predetermined set of questions, to highly unstructured, where the interviewer probes and explores the applicant's qualifications in a "play-by-ear" fashion (p. 145). Of course, many variations exist within and between these two extremes. Guion (1998), for example, lists four different types of structured interviews (i.e., patterned interviews, behavior description interviews, situational interviews, and comprehensive structured interviews).

Many factors influence the outcome of interviews. Among these factors are (a) temporal placement of information (applicants are less likely to be selected when negative information is presented early in an interview); (b) interviewer stereotypes of idealized successful applicants (some interviewers rate qualified applicants unfavorably because they do not fit with a stereotype of an ideal candidate); and (c) quality of applicants immediately preceding an interview (ratings of average applicants are strongly influenced if they are preceded by highly qualified or poorly qualified applicants) (Schmitt, 1976). The amount of structure during an interview is probably the factor that influences the outcome of the interview the most. This is because the interview is a dynamic process; interviewers affect the behavior of applicants, and vice versa (Muchinsky, 1987, p. 146). Having a structured interview guide increases interinterviewer agreement (Schmitt, 1976).

Issues of consistency and reliability apply to interviews just as they do to ratings of written materials. Two sources of reliability are important to evaluate: *intra*interviewer reliability and *inter*interviewer reliability. Intrainterviewer reliability refers to the similarity of judgments made by the same interviewer over time and interinterviewer reliability refers to the similarity of judgments made by different interviewers about the same applicant. Most research suggests that intrainterviewer reliability is high and that interinterviewer reliability is variable. Interinterviewer reliability tends to be low when the purpose of the interview is to evaluate "soft" variables. Muchinsky (1987) cites personality as an example of a variable that might be rated discrepantly; interviewers disagree as to what the personality construct means and how it gets translated into successful performance.

Conclusion

In order for individualized review in college admissions to be fair, issues of consistency and reliability must be considered. Every institution should ensure that interrater reliability remains high and that an applicant's file receives the same evaluation regardless of who reads it. There are a number of ways to assess interrater reliability, including calculating the composite reliability of readers, computing the proportion of times that readers make consistent ratings, and evaluat-

ing reader severity (average scores by reader). Examining the values that result from these calculations will provide insight into the existing level of interrater reliability and can guide an institution in improving the fairness and consistency of the individualized review process. Practices such as holding regular calibration meetings, conducting reader training sessions, using meaningful rubrics in the rating process, or increasing the number of readers that review an application should increase reliability and consistency. Similar practices may also be useful in increasing the reliability and consistency of interviewer evaluations.

There is no "best practice" for making college admissions decisions, as each institution has its own mission and unique needs. It is also true that colleges and universities face different challenges in finding the most appropriate and effective ways to maintain sufficient interrater reliability in their ratings of college applications. Because reliability is essential for guaranteeing that each applicant receives a fair evaluation, the assessment of interrater reliability becomes a major consideration for the college admissions community.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

College Board. (2004). *A Guide to the New SAT® Essay*. New York: College Entrance Examination Board. Retrieved July 13, 2004, from http://www.collegeboard.com/newsat/hs/scoring/practice.html.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Prentice Hall.

Muchinsky, P. M. (1987). *Psychology applied to work: An introduction to industrial and organizational psychology*. Chicago, IL: The Dorsey Press.

Parke, C. (2001). An approach that examines sources of misfit to improve performance assessment items and rubrics. *Educational Assessment*, *7*(*3*), 201–225.

Rigol, G. W. (2003). *Admissions decision-making models: How U.S. institutions of higher education select undergraduate students*. New York: College Entrance Examination Board.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Rubin, A., & Babbie, E. (1993). *Research methods for social work* (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Schafer, W. D., Swanson, G., Bené, N., & Newberry, G. (2001). Effects of teacher knowledge on rubrics on student achievement in four content areas. *Applied Measurement in Education*, *14* (*2*), 151–170.

Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology*, *29*, 79–101.

Vogt, W. P. (1999). *Dictionary of statistics and methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(*2*), 263–287.

Office of Research and Psychometrics
The College Board
45 Columbus Avenue
New York, NY 10023-6992
212 713-8000