Adjusting for Rater Severity Over Time

Thomas R. O'Neill

Performance assessments are often thought to have greater validity than multiple-choice tests because the rated behavior more closely approximates the behavioral domain of interest than does merely asking questions about it. For example, expert judges rating a student on a karate test want to know if the student knows how to strike with penetrating force. It seems obvious that asking the student to break a brick is more informative than asking the student to answer questions about breaking a brick. While breaking a brick is usually unambiguous with regard to success, other activities such as judging technique must be graded by raters. Because raters often have different individual standards of excellence, the reproducibility of estimates derived from ratings is sometimes questioned. Any given rating will be influenced not only by the examinee's ability and the task's difficulty, but also by a third facet, rater severity.

In order for measurements to be meaningful, differences in raters must be accounted for, so that all results are expressed from the same frame of reference. The extension of the Rasch (1960/1980) model to the Many Facet Rasch Model (MFRM, Linacre, 1989) has made accounting for rater severity possible by placing rater severity in the same frame of reference as item difficulty and examinee ability. The MFRM estimates each rater's severity, each project's difficulty, and/or other such facets, and removes their influence before computing an examinee's ability. An examinee's measure is independent of which rater graded them and which tasks they performed.
An alleged drawback is that with each additional link required to connect a test form back to its original scale, more error accumulates. However, it is often overlooked that with each successive administration, more historical data is available to guide the test development process.

A linking strategy is usually employed to align the scale defined by the current test administration with the original scale, thus the same scale is maintained across several administrations. In multiple-choice test, this linking is usually accomplished using several items common to both the current test form and the preexisting scale. Once the difficulty of the items on the new are aligned with the preexisting scale, an examinee taking two forms of the test will receive a comparable measure even when the test forms are different in difficulty.

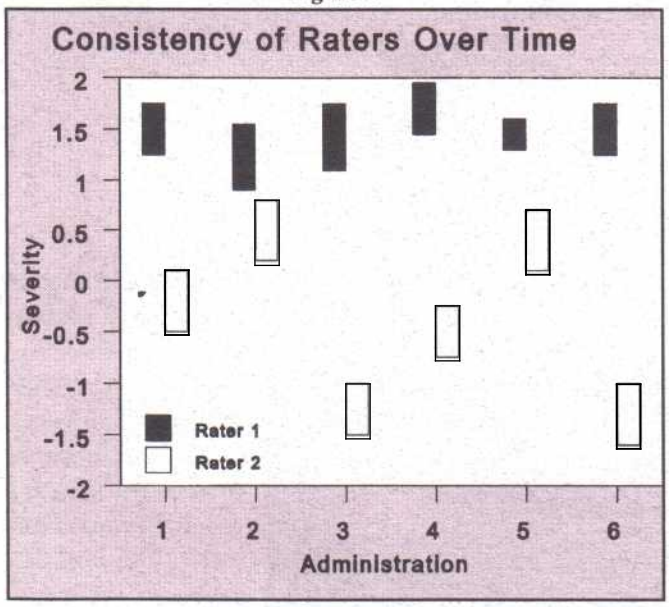
In performance assessments, the difficulty of the prompts from the current form must be aligned with the prompt's preexisting scale, but this alone is not enough to make the examinee measures comparable. The severity of the raters must also be aligned. It is important that the severity of new raters is expressed in the same frame of reference as that of the original raters. Using common raters to link together two test administrations requires that the common raters maintain a uniform degree of severity across administrations. However, actual raters occasionally violate this requirement and thereby potentially thwart our intention to carry forward the same scale. For this reason, it is important to use historical data to identify those raters who are most likely to maintain a uniform degree of severity across administrations.

As part of the equating process, rater stability is verified from administration to administration. This is done by comparing the severity of several common "anchor" raters on the current administration with their degree of severity from the prior administration, and then checking that their severity on the current administration places them in the same relative position as in the past. When their relative positions hold, it is reasonable to conclude that their severity has not changed. In cases where only one or two of the anchor raters have changed positions, it is reasonable to conclude that those one or two raters have changed their degree of severity and should be treated as new raters, but the rest of the anchor raters can be used to link the new raters to the established scale. But when several raters change places and the number of anchor raters is few, it becomes more complicated to determine which of the anchor raters changed their severity and which remained the same.

To prevent this from happening, psychometricians try to employ as many stable pre-calibrated raters as possible, so that any anomalous raters will stand out more clearly. While it can never be known in advance exactly how severe a particular rater will be on any given occasion, a rater's past performance can suggest how severe they will be in the future. Thus, historical information can be helpful to psychometricians who are organizing or equating performance assessments across administrations. By plotting a rater's severity with their error bands (±2 SEs) across administrations (**Figure 1**), psychometricians can verify that things are going well or identify problem areas. A method to do this can be found in Objective Measurement: Theory into Practice (volume 5).

Analyzing rater severity overtime should be part of the ongoing equating procedure because it can aid developers with historical data in making decisions about raters. For example, a psychometrician may select a few raters to participate in several consecutive administrations for the purpose of maintaining the same frame of reference for rater severity. Common raters should be selected on the basis of their documented ability to maintain a uniform level of severity. Armed with historical information **(Figure 1)**, psychometricians can seek out

Figure 1

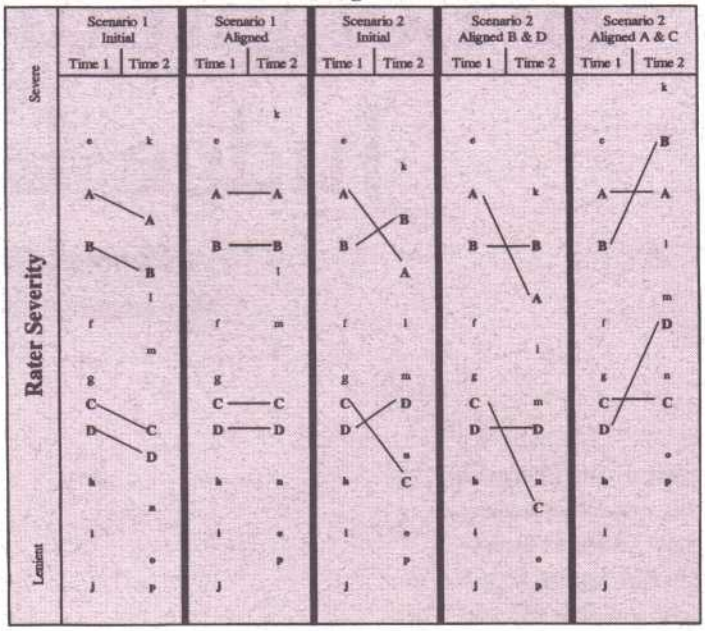


stable raters like Rater 1 for this purpose. Others, like Rater 2, can still be used across administrations because their degree of severity is consistent within administrations, but knowing their across-administration degree of severity has more variance, the psychometrician would not want to use them as a link back to the initial scale. Rater 2 should be thought of as a new rater each time he grades.

Viewing rater severity in this manner can generate hypotheses regarding how individual raters behave over time. When the psychometrician thinks that there has been a shift in a rater's severity and that the new level severity is likely to be stable, the psychometrician should update the rater calibration bank with the new severity calibration.

The most obvious information noticeable from **Figure 1** is which raters are consistent and which are erratic across administrations. This information can be used to select anchor raters, but it can also be used after the data has been collected. Suppose that out of ten raters, only four raters, A, B, C, and D had a known degree of severity **(Figure 2)** established from earlier administrations. Ideally, one would hope for results similar to the second administration as found in scenario 1 (initial). Because the four raters maintained their relative position from each other, aligning the common raters is a simple matter (scenario 1, aligned) which allows the severity of raters K through

Figure 2



P to be expressed in the same frame of reference as raters A through J.

However, suppose that two of the common raters changed their severity by approximately the same amount on the second administration as represented in scenario 2 (initial). How would the psychometrician know if A and C became more lenient (scenario 2, B & D aligned) or if B and D became more severe (scenario 2, A & C aligned)? Either scenario seems equally plausible. A potential answer is to review the historical performance of the four raters. It seems probable that the historically more stable raters would be less likely to be the ones who changed.

To prevent the above scenario, enough common raters should be employed so that if a small percentage of raters change in severity, it will be easy to identify which raters changed. Reviewing the historical data can allow the psychometrician to make a good guess that, given the available pool of stable, pre-calibrated raters, (1) which raters should be selected, (2) how many of the raters are expected to change severity during this administration, and (3) how many raters will be needed to clearly identify those who have changed severity.

Thomas R. O'Neill is Manager of Research & Analysis for the American Society of Clinical Pathologists. His professional interests include performance assessment, Computerized Adaptive Testing (CAT), and the promotion and creation of understandable and useable measurement. His personal interests include Korean Tang Soo Do (karate), zymurgy (beer brewing), and travel.

email: tom_o'neill@ascp.org (the apostrophe is important!)