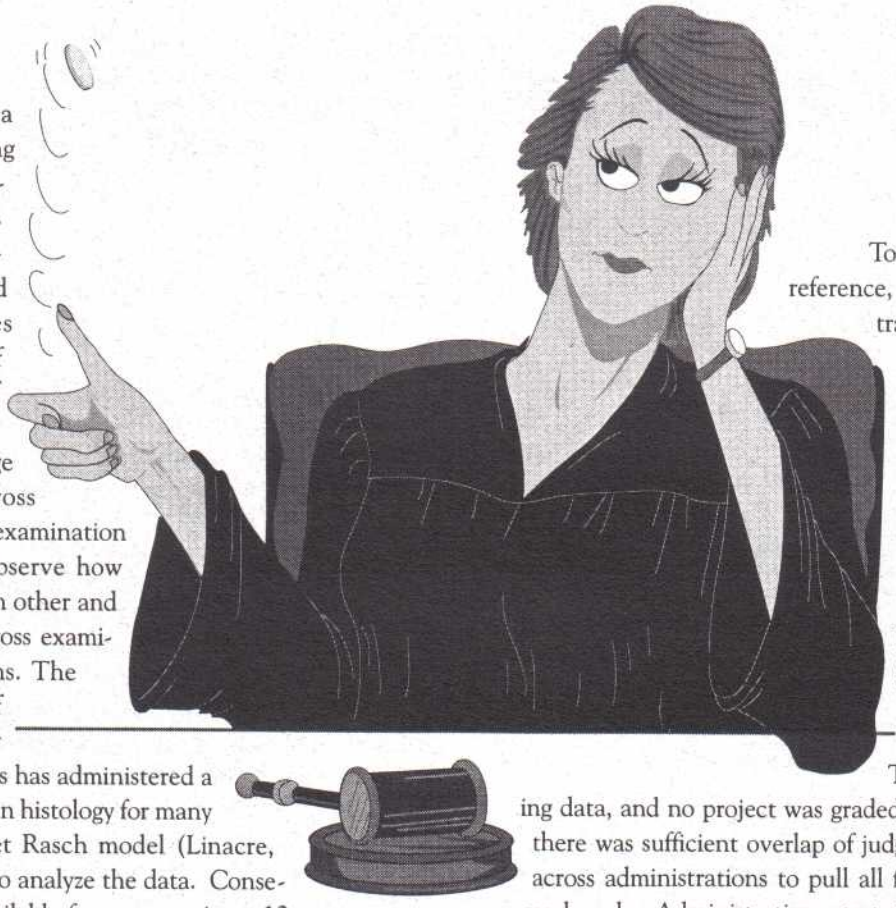
A Longitudinal Study of Judge Leniency

Mary E. Lunz, Ph.D.
Measurement Research Associates, Inc.

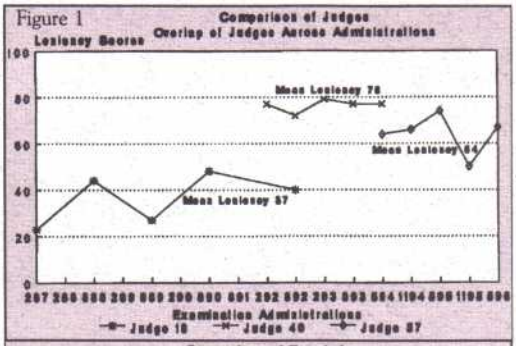


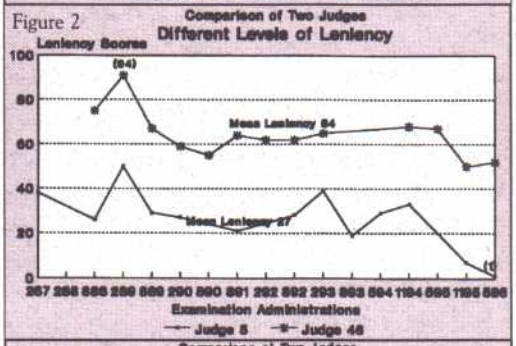
The judge is a critical part of scoring any clinical examination. However, little is known about long-term consistency and leniency. Do judges change their level of leniency over time; if so, in what directions? This study tracks judge grading patterns across ten years of clinical examination administrations to observe how judges differ from each other and within themselves across examination administrations. The Board of Registry of the American Society of Clinical Pathologists has administered a clinical examinations in histology for many years. The multi-facet Rasch model (Linacre, 1989) has been used to analyze the data. Consequently, data were available for constructing a 10-year longitudinal study of judge performance. The clinical examination has four facets: 1) candidates, 2) judges, 3) projects, and 4) tasks. Over the ten years there were 4,683 candidates, 57 judges, and 53 projects. Three tasks were graded at each administration. Two were graded as 1=acceptable and 0=unacceptable and the third task was graded on a four-point scale as 3=excellent, 2=acceptable, 1=marginal, 0=unsatisfactory. The same grading scales were used for all administrations. Candidate performances were randomly assigned to judges. Each candidate was judged on the three tasks for 15 projects, with input from three judges. All judges graded examples of all projects during each administration.

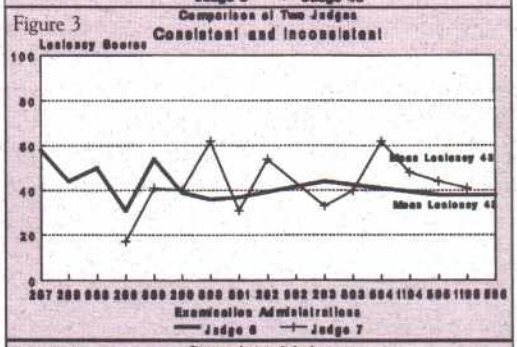
To construct a frame of reference, data from 17 administrations were pooled and analyzed together. This placed all examination administrations for ten years on the same "benchmark" scale. The FACETS program (Linacre, 1994) was used to calibrate candidate ability, judge leniency, project difficulty, and task difficulty on this scale. There was a lot of missing data, and no project was graded more than once. But there was sufficient overlap of judges, projects, and tasks across administrations to pull all facets onto the benchmark scale. Administrations started in February, 1987 (labeled 287) and continued semi-annually through May, 1996 (596).

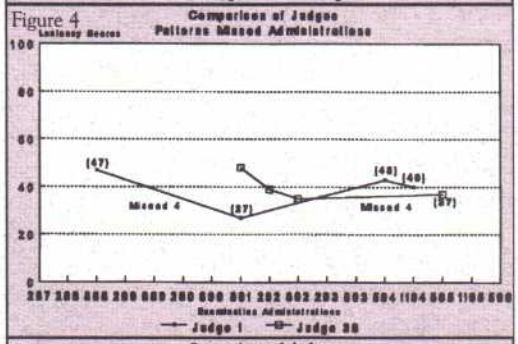
After the benchmark scale was constructed, individual examination administrations were re-analyzed separately. The difficulty estimates for the projects and the tasks, as well as the candidate ability measures from the benchmark scale, were used to anchor the individual examination administrations. The non-anchored facet across administrations was judge leniency. This enabled differences in judge leniency to be tracked across administrations. The multi-facet judge leniency estimates were transcribed to scaled scores so that 0 points marked the most lenient judge and 100 marked the most severe judge.

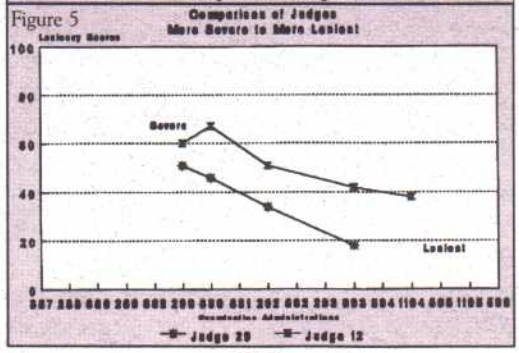
On average, judges graded in six administrations, but the range was 1-15 administrations. Different subsets of judges graded during each administration. However, there were always some judges that overlapped among administrations. **Figure 1** shows that judges 18 and 57 were linked with judge 40. Mean candidate ability estimates across administrations were verified as not significantly different among test administrations. Most judges graded in some administrations and skipped others. Some judges graded many sessions, while others graded few. Some judges varied among administrations, while others were extremely consistent. The graphs show examples of judge grading patterns across administrations. **Figure 2** shows the comparison of a relatively severe and a relatively lenient judge. The mean leniency of judge 46 was a scaled score of 64 points, while the mean leniency of judge 5 was a scaled score of 27 points. Each of these judges graded in 13 administrations and varied within 20 points of their average leniency across all examination administrations. **Figure 3** shows judges who are consistent and inconsistent in their leniency among administrations. Each of these judges graded at 10 of the 17 administrations. The average leniency of both of these judges was a scaled score of 43; however, judge 7 tended to vary in overall leniency at each administration, while judge 6 showed little variance after the first several examination administrations, even when administrations were missed. **Figure 4** shows that judges are consistent in their leniency even when they do not grade in consecutive examination administrations. Judge 38 graded three consecutive administrations, then missed four consecutive examination administrations, but stayed within a 10-point leniency range. Judge 1 graded in one administration, then missed four administrations, then graded one administration, then missed four administrations, but remained within a 10-point leniency range. **Figure 5** shows two judges who moved from relatively severe to relatively lenient. Some sessions were missed, but the pattern of becoming more lenient is obvious for these judges. The study shows that clinical examination data from different examination administrations can be placed on a benchmark scale when there are commonalties that link examination administrations using the multi-facet model. Some judges were consistent across years; however, some were less consistent, possibly because of limited grading experience, educational or personal changes, or technical experience.











Mary E. Lunz, Ph.D.

Mary E. Lunz earned a Ph.D. from Northwestern University. After teaching and consulting for several years, Mary worked as Director and Psychometrician for the Board of Registry of the American Society of Clinical Pathologists for 17 years. During this time, she began working with Ben Wright and Michael Linacre on issues relating to performance examinations, and computerized adaptive testing. Research is still ongoing.

Mary is currently Director and Senior Associate at Measurement Research Associates, Inc., which provides psychometric services to medical specialty, dental specialty, and allied health certification organizations. Her specialty is the development and analysis of performance examinations, especially oral interview examinations. (MeasResInc@aol.com)