Educopia

Brief Summary of RESA Design 2013-14

The RESA is a performance assessment that requires candidates to demonstrate their knowledge and skills in real time. The 2013-14 and 2014-15 RESA consisted of five tasks[1]:

- Tasks 1 and 3—Lesson Cycle #1 and #2, with 11 rubric components;
- Task 2—Formative and Summative Assessment, with 5 rubric components;
- Task 4—Communication and Professional Growth, with 5 rubric components; and
- Task 5—Reflection on Teaching Practice Based on Feedback from Students and/or Colleagues, with 1 rubric component

Each rubric component except Task 5 was scored on four levels; Task 5 responses were scored as compliant or non-compliant. Component scores were summed to create task-level scores, and the passing standards were set in this metric. The task scores were placed on the reporting scale using Rasch 1PL IRT modeling. Candidates are allowed to "bank" passing scores and only retake those RESA tasks that they did not pass. RESA Task 2 was significantly revised during 2013-14 and will not be directly comparable to subsequent Task 2 scores. Ohio decided to treat RESA Task 2 scores using a "completion" standard for 2014. Candidates who successfully submitted all Task 2 materials and received at least the minimum score for each rubric were judged to have passed the task. A new standard setting and cut score will be determined for this task in 2016.

Factor Structure

Analyses indicate that a bifactor model is the best fit to the RESA data for Tasks 1-4,[2] in which there is one general teacher performance factor and the task-specific skills and abilities that were assumed to be mutually independent. Tasks 1 and 3 both tend to have two roughly-defined factors within their data, one consisting of the video components and one written response that refers to the video, and another consisting of the remaining written responses. Both Task 2 and Task 4 form a single structural factor in the data analysis.

Validity Evidence

External validity was investigated using the relationship with OTES scores, for the subset of candidates who had both RESA and OTES scores. Only about 45% of the RESA candidates submitted Tasks 1-4 and had an OTES score that could be matched. For all RESA tasks, the correlation between the scores was weak (around 0.08-0.15) but always positive and statistically significant. This suggests the RESA and OTES measure distinct, but related, constructs.

Potential Bias Review

A review of the RESA handbook, assessor training materials, rubrics, and websites for biasing factors, sensitive language, or language that is potentially difficult to access for individuals or subgroups was completed. No indications of substantively biased language were discovered; there were some minor clarifications recommended. In the Differential Item Functioning (DIF) analyses, no item was flagged across all the subgroups. Two items were flagged in two comparisons and will be monitored in year 2 of RESA to see if the differences persist. In addition, descriptive comparisons of pass rates and average task scores were conducted. A statistical significance test

[1] For the 2015-16 RESA, Task 5 was eliminated for first-time candidates.
[2] Task 5 is not included in these analyses due to the two-level scoring.

of differences in pass rates was not done, as these tests are highly dependent upon sample size and the groups of interest in RESA vary considerably in size. There are observed differences in the percentage of subgroups in passing rates on RESA tasks among subgroups: Female candidates tend to have higher pass rates than male candidates; non-minority candidates tended to have higher pass rates than minority candidates; candidates with a graduate degree tended to have slightly higher pass rates than those with undergraduate degrees; and candidates who work in urban or suburban schools, in schools with average or low levels of student poverty, or in districts with small student populations, tended to have higher pass rates than their counterparts in rural or small town schools, in schools with high levels of student poverty or in districts with larger student populations. In all group comparisons, the full score distributions cover very similar ranges, and for all tasks, the 25th percentile is at or above the cut score for all groups.

Standard Setting

The passing standards for the tasks were set by a group of Ohio educators and stakeholders. Their roles at the time of the meeting included elementary school teacher, middle school teacher, high school teacher, special needs teacher, career and technical education (CTE) teacher, principal, administrator, program coordinator, and representative from institutions of higher education (IHEs). The purpose of the standard setting was to set a cut score that would indicate a “just sufficiently qualified” candidate. This is the lowest performance on the RESA task that would signal readiness for professional licensure. Panelists used impact data (the percentages of candidates that would be categorized as passing using those cut scores) to revise their initial cut score recommendations and determine the final cut cores used operationally.

Scoring

RESA tasks are each scored on a single set of rubric components, regardless of the grade level or content area of the submission. In order to minimize error, raters are trained and assessed for scoring accuracy before being allowed to score live submissions. Between 70% and 80% of those invited to training eventually certified to score Tasks 1-4; Task 5 had a 100% certification rate. In addition to certification, rater accuracy was assessed through calibration (pre-scored submissions seeded through all raters’ scoring queues). Raters must pass certification in order to continue scoring. Task 2 had the lowest calibration success rate at approximately 66%; raters calibrated successfully on Tasks 1, 3, and 4 at between 80% and 85%. In addition to calibration assessments, where raters are given direct feedback on their success, there are validity cases. These are also pre-scored, but the raters are not given feedback on their accuracy in scoring these submissions.

RESA scoring is unusual in that all quality monitoring is done against “correct” scores, not against other rater scores. No double-scoring is done for monitoring purposes, although standard double-scoring agreement statistics can be calculated using the data from the validity cases. Kappa, a statistic that corrects for matching that can occur by chance, was used to evaluate rater agreement on RESA. Generally raters exhibited fair levels of agreement on Tasks 1, 3, and 4, and slight agreement on Task 2 as kappa is typically interpreted.

Score Reliability

Two estimates of reliability were calculated separately for each task. Tasks 1-4 all demonstrate adequate levels of reliability, between 0.75 and 0.90. Perhaps of greater interest are the estimates of error around the scores at which pass/fail decisions are made. These errors may be used to estimate the accuracy of candidate classification; that is, how often does the observed score correctly assign to a candidate the status of “passed” or “failed”. Tasks 1-4 have classification accuracy at or above 95% each.