Developing Reliable Student Assessments

Yale Center for Teaching and Learning
https://ctl.yale.edu / https://ctl.yale.edu/ReliableAssessments

Reliability refers to how well a score represents an individual's ability, and within education, ensures that assessments accurately measure student knowledge. Because reliability refers specifically to score, a full test or rubric cannot be described as reliable or unreliable. Rather, reliable scores help students grasp their level of development, and help instructors improve their teaching effectiveness. A variety of methods are commonly used to estimate reliability of scores, and instructors can make reliability methods transparent to motivate student effort and assure them of accuracy.

Instructors should note that there are many reasons why a score may not perfectly represent a student's knowledge. For instance, test anxiety, distractions in the testing environment, or guesswork could cause discrepancies between a score and an individual's actual ability. While some of these factors cannot be completely eliminated, instructors can improve reliability when designing assessments, grading student work, and analyzing student performance on individual test items or criteria.

Recommendations

Reliability can be increased by several methods. If the evaluation is performance- based or an essay:

- **Design a rubric** – Rubrics help the evaluator(s) / grader(s) focus on the same criteria across all submissions. **Rubrics** can be designed in a variety of ways, and also make grading standards and performance expectations clear for students.

- **Grade item by item** – If students are given multiple essays or problem sets, instructors can evaluate/grade the first essay/problem on each student's paper before grading the second essay/problem. This allows the evaluator/grader to apply the same set of criteria at a time, and minimizes the effect of the impact of fatigue or mood differentially affecting any one student's performance.

- **Grade anonymously** – Instructors may wish to know whose work they grade, to provide feedback about course-wide performance. However, every grader/evaluator possesses some biases, which can either positively or negatively affect individual students score. For instance, if a student is a hard worker in class, an instructor may be more lenient when grading an essay from that student. Instructors can grade anonymously to minimize the effect of bias in the grading process. Instructors can bypass a student's name when grading, or consider other **blind grading** approaches.

- **Train graders** – If multiple graders are being used, instructors should provide training to the graders on how to utilize rubrics or evaluation/grading criteria. **Sample essays** or performance can be provided. Additionally, for each essay or problem, a subset of submissions should be independently scored by multiple graders. Inter-rater reliability can be calculated on the subset, and the graders can discuss any discrepancies before grading the rest of the submissions.

If the evaluation consists of a multiple-choice test or Likert-type items:

- **Design the assessment using a table of specifications -** A table of specifications outlines the content that is covered in a test or assessment. A table of specifications typically consists of three main components. *First,* a list of topics that are covered on the assessment. *Second,* a classification or taxonomy (i.e. **Bloom's taxonomy**) that describes the types of questions that are on the exam. *Third*, an indicator of the number of questions to be presented that corresponds to each content area and classification.

Topic or Content Area	**Multiple choice questions measuring recall**	**Multiple choice questions measuring application**	**Multiple choice questions measuring evaluation**	**Total Number of Questions**
Chemical Reactions	Q 1, 6, 7	Q 12, 14, 17, 19	Q 21, 24, 26, 29, 30, 35, 38, 39	15
Thermodynamics	Q 2, 3, 8, 9	Q 11, 15, 18	Q 22, 25, 31	10
Chemical Equilibrium	Q 4, 5, 10	Q 13, 16, 20	Q 23, 27, 28, 32, 33, 34, 36, 37, 40	15
Total Number of Questions	10	10	20	40

Sample Table of Specifications: Using Components of Bloom's Taxonomy

- The table of specifications allows for subscales to be created among multiple concepts being tested. For instance, separate reliability coefficients can be calculated for items that test the first unit and items that measure the second unit. A table of specifications will also provide detailed feedback to students and instructor about content covered.

- **Conduct item-level diagnostics to improve the test. Please note that some testing software can provide the data described below for you in the form of a report.**

    - **Cronbach's alpha** – When calculating Cronbach's Alpha, it is possible to determine which items are negatively impacting reliability. Those items could then be removed to increase the reliability of the score.

    - **Item difficulty** – The percentage of students who answered an item correctly. Items that are too difficult can negatively impact reliability, if difficulty can successfully be related to the question or content, and not to student study performance. However,

items that are too easy do not detect differences between high and lower performing students.

- **Item discrimination** – Examines how well an item can discriminate between high performing and low performing students. Items that do not perform as expected (higher performing students get the answer right more than lower performing students) negatively impact reliability.

- **Distractor analysis** – Determines which distractor questions students (or students of different performance levels) choose. Any distractor that is not selected (or is rarely selected) should be changed. If students can eliminate answer choices, they have a higher probability of guessing the correct answer without understanding the content.

Further Reading

Cronbach LJ. (1951). Coefficient alpha and the internal structure of tests. Psychometrika 16: 297-334.

Guttman L. (1945). A basis for analyzing test-retest reliability. Psychometrika 10: 255-282.

Gwet KL. (2014). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC.

Malouff J. (2008). Bias in grading. College Teaching 56(3):191-192.

Murphy KR & Davidshofer CO. (1988). Psychological testing. Principles, and Applications. Prentice Hall: Englewood Cliffs, NJ.

Osterlind, SJ. (2006). Modern measurement: Theory, principles, and applications of mental appraisal. Pearson: Upper Saddle River, NJ.