Research & Development

Listening. Learning. Leading.®

Chapter 2
Reliability, Precision, & Errors of Measurement

AERA, San Antonio, 2017

Michael Kane

Classical "Reliability" and Precision

- Reliability Coefficient: the ratio of true-score variance to observed score variance.
- Reliability/precision (R/P): the consistency of scores, interpretations, and decisions across replications of the measurement procedure (MP).
- Analyses of R/P depend on the kinds of variability allowed by the MP and the intended interpretations and uses of the scores.

For example,

- if the score interpretation assumes invariance over time, then variability over occasions is considered error.
- If scores from different test forms are considered exchangeable, then variability over forms is considered error.
- Qualified raters are considered exchangeable, so variability over such raters is considered error.
- For state variables, which can vary over occasions, variability over occasions is not error.

SEMs
Standard Errors of Measurement

- SEMs are conceptualized in terms of the standard deviation of the errors over replications of the MP.
- SEMs cannot typically be directly estimated for individuals.
- So, we estimate the SEM indirectly by estimating some R/P index related to the SEM.
- We may then interpret R/P in terms of the index, or use it to estimate an average SEM or a conditional SEM.

"True scores" and "Errors"

- To say that a score has error implies that there is some error-free value of the variable.
- In classical test theory, we have true scores and error
- In G-theory, we have universe scores and multiple sources of error.
- In IRT, we have theta values and errors.

Reliability and Validity

- Reliability is a necessary but not sufficient condition for validity.
- It evaluates the extend to which scores can be generalized over various conditions of observation (occasions, tasks, contexts, raters)
- And the extent of generalization is a basic part of an interpretation.

Replications

- Replications are independent administrations of the MP, and variability over replications is considered *random error*:
  - Alternate forms (or parallel forms)
  - Internal consistency (e.g., coefficient alpha)
  - Test-retest
  - Interrater

Evaluating Reliability/Precision

- It is important to recognize the sources of error in any coefficient or SEM (classical, G-theory, IRT)
- G-theory tends to be most explicit about this, because it specifies variance components to be included in the error.
- If we want to make claims about overall error or precision, we should include all sources of error in the analysis.

Factors Affecting R/P

- Population
- Test length (number of independent tasks)
- Nature of tasks and scoring
- Training of raters.
- Numbers of raters for each task
- Construct Definition – facets included in error
- Definition of MP – fixed and random facets

SEM
Standard Error of Measurement

- SEMs can be used to create confidence intervals around point estimates.
- Like reliability and G coefficients, SEMs can include one or more sources of error, and their values depend on population, test length, definition of the construct, etc.
- SEMS tend to be most useful when they include all relevant sources of error.
- Otherwise, they tend to be underestimates of the error.

Decision Consistency

- In cases where the main use of scores is to assign students to categories (e,g, basic, proficient, etc., or pass/fail) indices of decision consistence are particularly relevant in evaluating precision.
- Decision consistency can be improved by reducing SEMs around cut scores.

Estimates of Group Means

- In estimating group means, the sampling of individuals is a source of error, if the sampling can be considered random.
- For example, the students in a class or grade level in a school or district could have been different and will be different next year.
- As a result, group means based a MP that tends to produce reliable/precise scores may not yield precise estimates of group means.

Documentation 1

- Reported indices of R/P should identify the sources of error in the scores, given the proposed interpretation/use of the scores.
- In reporting indices of R/P, the methods used to collect data and to estimate the index should be made clear.
- If scores are used to classify students, decision consistency should be reported.

Documentation 2

- If group means are reported, the R/P of these mean scores should be reported.
- If the scores reported and used are more complex functions of student test scores and other variables (e.g., VAMs), the R/P of these reported scores should be reported.

Standards for Reliability/Precision

General Standard
Standard 2.0

- Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use.

8 Clusters of Standards

- C1: Specifications for Replications of the Testing Procedure
- C2: Evaluating Reliability/Precision
- C3: Reliability/Generalizability Coefficients
- C4: Factors Affecting Reliability/Precision
- C5: Standard Errors of Measurement
- C6: Decision Consistency
- C7: Reliability/Precision of Group Means
- C8: Documenting Reliability/Precision

C1: Specifications for Replications of the Testing Procedure

- 2.1 State the range of replications over which R/P is being evaluated, along with a rationale for this definition, given the testing situation.

- 2.2 The evidence for R/P should be consistent with the domain of replications, and with the intended interpretations and uses of scores.

C2: Evaluating Reliability/Precision

- 2.3 For each score or combination of scores that is interpreted, relevant R/P evidence should be reported.
- 2.4 When the interpretation emphasizes differences between scores, R/P data should focus on such differences.
- 2.5 R/P evidence should be consistent with the structure of the test.

C3: Reliability/Generalizability Coefficients

- 2.6 Indices that address one kind of “error” should not be considered interchangeable with indices that address other kinds of “error”.
- 2.7 When performances are scored subjectively, evidence of rater consistency should be provided in addition to any other relevant kinds of R/P data.

C4: Factors Affecting Reliability/Precision

- 2.8 If responses are scored locally, R/P should also be evaluated locally.
- 2.9 If long and short forms are available, R/P should be evaluated for both.
- 2.10 If variations in testing are permitted, R/P would be evaluated for each variation.
- 2.11 Provide R/P evidence for subgroups.
- 2.12 if separate norms are provided for age groups, R/P should be evaluated for each age group.

C5 Standard Errors of Measurement

- 2.13 The SEMs should be in units of score or subscore scales.
- 2.14 If possible, conditional SEMs should be reported at several levels, particularly at cutscores, if relevant.
- 2.15 Any indications that conditional SEMs might differ substantially across subgroups should be investigated.

C6: Decision Consistency

- 2.16: if scores are to be used to make classification decisions, the percentage of test takers classified consistently across replications should be reported.

C7: Reliability/Precision of Group Means

- 2.17: When average scores for groups are reported, R/P evidence should reflect sampling of examinees, as well as individual errors.
- 2.18: When complex sampling schemes (e.g., matrix sampling) are used, R/P analyses should reflect the sampling scheme.

C8: Documenting Reliability/Precision

- 2.19: Methods used to estimate R/P indices should be described clearly, and the sampling of test takers in the analyses should be reported.
- 2.20: If R/P indices are adjusted for restriction of range, supporting rationale, descriptive statistics, and both adjusted and unadjusted results should be reported.

2 Scenarios

Scenario 1

- A licensure test is used to admit candidates to professional practice. A new form of the test is administered on each testing date. Each form of the test includes an objective test and a performance test. The two subtests are each equated across administrations, and the sum of the two scaled scores is used to make pass fail decisions, based on a predefined cut score.
- Aggregate results for states are also reported.
- How should we evaluate the reliability/precision of this testing program?

Scenario 2a

- A state test is administered all students at a grade level in the state. A raw score and an equated, scaled score are generated. Only the scaled scores are reported to the students, their parents, their schools and their teachers.
- Student scores are also classified into 4 categories (below basic, basic, etc.). These results are also to be reported to students, parents, schools, and teachers
- The numbers of students in each category are reported for the state, school districts, and schools.

Scenario 2b

- What kinds of reliability/precision evidence would be appropriate for this testing program?
- What if responses to the performance tasks are scored locally?
- What if the scaled scores are also to be used for value-added evaluations of schools and teachers?

Thank you