ITEMS
Instructional Topics in Educational Measurement Series

NCME national council on measurement in education

"ITEMS Corner" Module Presentation: Reliability in Classical Test Theory

Charlie Lewis, Educational Testing Service & Michael Chajewski, Kaplan Test Prep

Module Overview

In this ITEMS module we provide a two-part introduction to the topic of reliability from the perspective of *classical test theory* (CTT). In the first part, which is directed primarily at beginning learners, we review and build on the content presented in the original didactic ITEMS article by Traub & Rowley (1991). Specifically, we discuss the notion of reliability as an intuitive everyday concept to lay the foundation for its formalization as a reliability coefficient via the basic CTT model. We then walk through the step-by-step computation of key reliability indices and discuss the data-collection conditions under which each is most suitable. In the second part, which is directed primarily at intermediary learners, we present a distribution-centered perspective on the same content. We discuss the associated assumptions of various CTT models ranging from parallel to congeneric, and review how these affect the choice of reliability statistics. Throughout the module, we use a customized Excel workbook with sample data and basic data manipulation functionalities to illustrate the computation of individual statistics and to allow for structured independent exploration. In addition, we provide quiz questions with diagnostic feedback as well as short videos that walk through sample exercises within the workbook.

Key words: reliability; classical test theory; KR-20; KR-21; Cronbach's α; Pearson correlation; Spearman-Brown formula; parallel model; tauequivalent model; congeneric model

Prerequisite Knowledge

This ITEMS module assumes no a priori technical knowledge of the concept of reliability. However, it is probably helpful to have a working knowledge of foundational assessment and statistical concepts such as:

- assessments / tests, items / tasks, and total scores / scale scores
- variable types, scales, and distributions
- basic summary statistics (mean, variance, standard deviation, correlation)

Specific prior experience with software is not required but access to Excel is necessary to use the interactive workbook. For illustration purposes a few basic Excel formula functions are used in sample solution videos; however, exploratory data manipulation in the workbook is done without these formulas through a user-friendly setup.

Learning Objectives

Upon completion of this ITEMS module, learners should be able to:

A. Beginner Section

- understand the intuitive foundations behind the concept of reliability
- understand how the concept of reliability is related to data and statistics
- understand how different assessment design factors affect the reliability of test scores
- understand the basic structure and variance components of the basic CTT model
- understand the formalization of the reliability coefficient in the basic CTT model
- understand key properties of the reliability coefficient
- differentiate between different data-collection designs and how they relate to the computation of different reliability statistics
- compute basic reliability statistics (KR-20, KR-21, Rulon, Cronbach's α, Pearson correlation, Spearman-Brown projection) from data with a calculator or Excel

B. Intermediate Section

- understand the distribution-based framing of reliability
- visually identify key marginal and conditional distributions in plots
- express univariate, bivariate, and multivariate distributions mathematically
- understand the derivation of basic reliability statistics (Rulon, Cronbach's α, Spearman-Brown projection, disattenuation formula) using distributional moments
- list the assumptions underlying the parallel, essentially parallel, tau-equivalent, essentially tau-equivalent, and congeneric measurement models
- understand which reliability statistic is appropriate for which measurement model(s)
- generate data with desired statistical properties and evaluate the impact of those properties on the values of different reliability statistics

After completion of this module, learners might wish to take additional ITEMS modules on reliability from the perspective of generalizability theory, item response theory, structural equation modeling, or diagnostic measurement as well as foundational models on these measurement frameworks. Check out the NCME or ITEMS Portal webpages for up-to-date information on available ITEMS modules!

Module Structure

The module is divided into the following sections, which can be reviewed sequentially or independently [*approximate completion times in parentheses*].

- Module Introduction [*5 Minutes*]
- Interactive Workbook [*30 Minutes*]

Beginner Section

- Section 1: Foundations [*15 Minutes*]
- Section 2: Formalizations (I) [*10 Minutes*]
- Section 3: Statistics (I) [*45 Minutes*]

Intermediate Section

- Section 4: Formalization (II) [*20 Minutes*]
- Section 5: Statistics (II) [*20 Minutes*]
- Section 6: Quizzes [*15 Minutes*]

Module Components

This ITEMS module includes the following components, which are delivered within a unified design shell created with modern course development software:

- integrated content slides that provide a structured walk-through of the content and computational examples with suitable voice-over and/or video components;
- an interactive Excel workbook with simulated sample data modeled after the Traub and Rowley (1991) didactic article as well as a data manipulation component that is used for worked examples throughout the module;
- text-based resources such as the original Traub & Rowley (1991) didactic overview article that reviews the key concepts in the beginner section as well as a new companion overview article that reviews the key concepts in the intermediate section;
- self-assessment questions with diagnostic feedback covering the conceptual, computational, and modeling aspects of the module;
- performance-based exercises that require data manipulation and interpretation with video walk-throughs;
- a glossary of key terms used in the module;
- a small library of online resources that have been vetted for general scientific accuracy and instructional utility.

Additional materials may be added over time so check back periodically!

Module Developers

Charles Lewis is a Distinguished Presidential Appointee at the Educational Testing Service (ETS) and Professor Emeritus of Psychology and Psychometrics at Fordham University. He also taught psychology and psychometrics at Dartmouth College, the University of Illinois, and the University of Groningen. His research interests include fairness and validity in educational testing; mental test theory, including item response theory and computerized adaptive testing; Bayesian inference; generalized linear models; and behavioral decision making. He was recently co-editor and co-author of *Computerized Multistage Testing: Theory and Applications* (2014).

Michael Chajewski received his undergraduate degree in experimental psychology from the University of South Carolina, and a masters degree in forensic psychology from John Jay College of Criminal Justice, The City University of New York. He received his doctoral degree in Psychometrics and Quantitative Psychology from Fordham University. As a psychometrician, Michael worked for eight years for the College Board supporting operational testing programs such as PSAT/NMSQT and AP, as well as assisted in the redesign of the SAT. His contributions and research spanned a variety of technical work including equating, test security and system development. Since 2017 Michael has been leading the psychometrics team at Kaplan Test Prep,

spearheading measurement model development for formative assessment and innovating assessment operating procedures. As an educator, Michael has taught both undergraduate and graduate courses within the CUNY system as well as at Fordham University. His research interests include configuring adaptive assessments, large data model fit evaluations, missing data impact, scaling, norming, as well as statistical software development and Bayesian statistics.

Associate Editor / Lead Instructional Designer

André A. Rupp is a Research Director at the Educational Testing Service (ETS) in Princeton, New Jersey. He is the co-author and co-editor of two award-winning interdisciplinary books entitled *Diagnostic Measurement: Theory, Methods, and Applications* (2010) and *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (2016). His research synthesis- and framework-oriented work has appeared in a wide variety of prestigious peer-reviewed journals. Among other things, he is passionate about improving processes for interdisciplinary collaborations during the development and implementation of scoring solutions for digitally-delivered assessments. Consequently, he is very excited to serve a three-year term as the associate editor / lead instructional designer of the emerging ITEMS portal for NCME that will provide various digital resources for members to support self-directed learning and professional development.

This is the pre-peer reviewed version of the following article:

Lewis, C., Chajewski, M., & Rupp, A. A. (2018). Digital ITEMS module 01: Reliability in classical test theory. Educational Measurement: Issues and Practice, 37(2), 71-72.

It has been published in final form at https://doi.org/10.1111/emip.12209. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.
