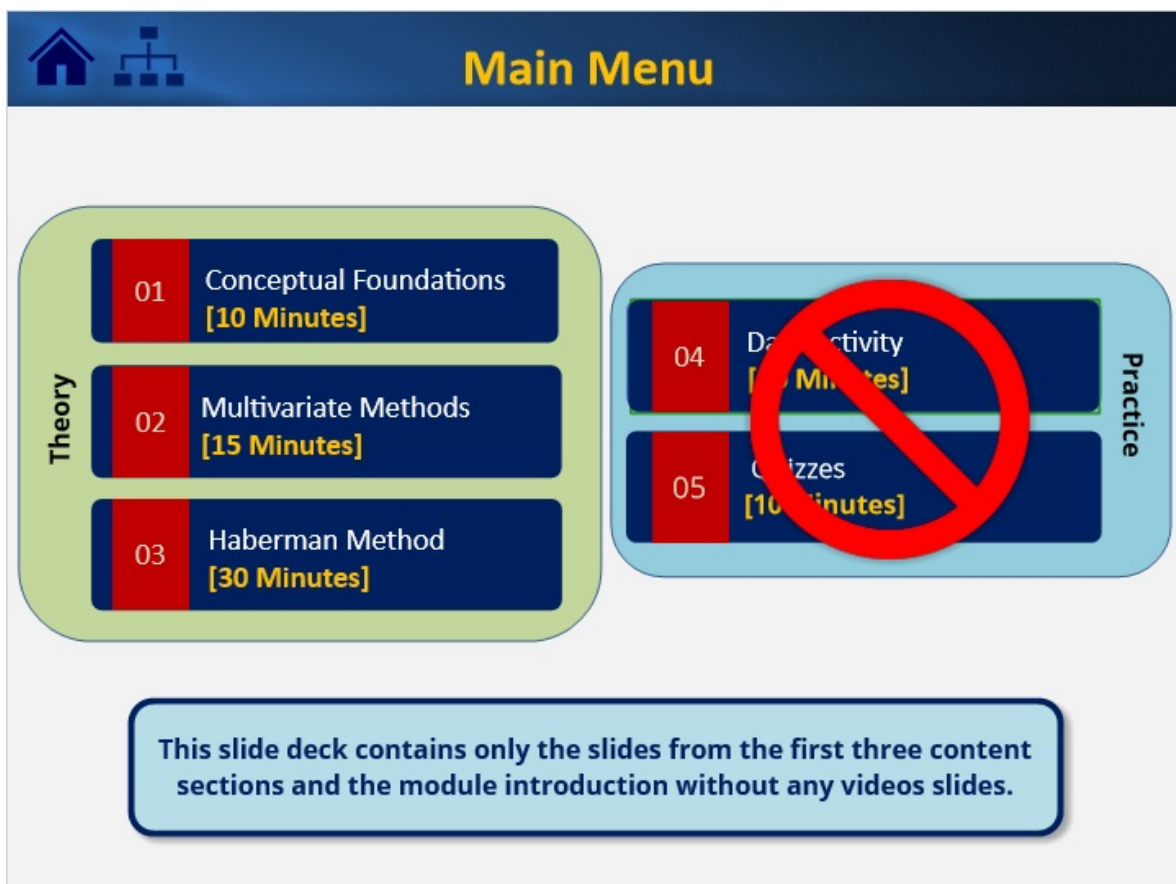


ITEMS Digital Module 07: Subscores – Evaluation & Reporting

This document contains all core content slides from sections 1-3 with the exception of slides that show video screens. In the digital module all slides can be accessed individually.

Module Organization

The module starts with an introductory section that leads to the main menu from which learners can select individual content and activity sections:



The screenshot shows a 'Main Menu' interface with a dark blue header containing a home icon, a tree icon, and the text 'Main Menu'. Below the header, there are two main sections: 'Theory' (left, green border) and 'Practice' (right, blue border). The 'Theory' section contains three items: 01 Conceptual Foundations [10 Minutes], 02 Multivariate Methods [15 Minutes], and 03 Haberman Method [30 Minutes]. The 'Practice' section contains two items: 04 Data Activity [10 Minutes] and 05 Quizzes [10 Minutes]. A large red 'X' is overlaid on the 'Practice' section. At the bottom, a light blue box contains the text: 'This slide deck contains only the slides from the first three content sections and the module introduction without any videos slides.'

1. Module Overview

1.1 Module Cover (START)





1.2 Author



1.3 Designers

Meet the instructional design team:






André A. Rupp
Educational
Testing Service

Xi Lu
Florida State
University

Click on the images to get to know them a bit!

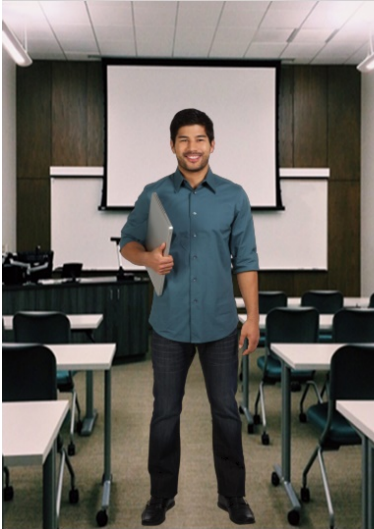
Additional
Thanks

Thanks (Slide Layer)



Back

1.4 Welcome



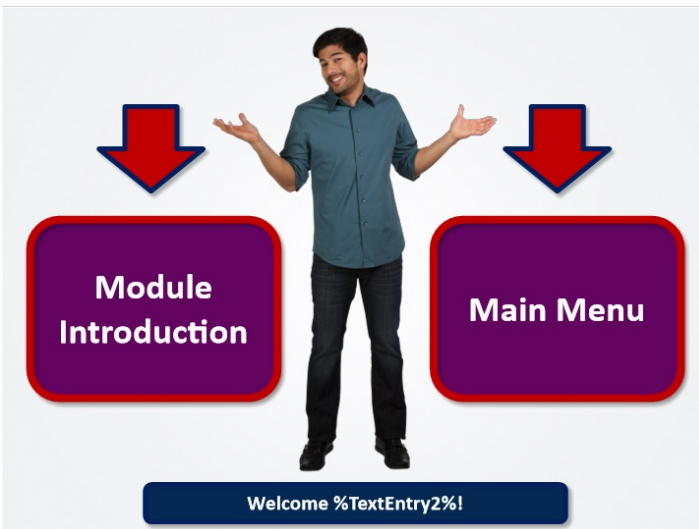
Welcome to the ITEMS Module!

The man to the left is Jet!

Along with the content developer he will be guiding you through the module content.

Tell us your name here:

1.5 Path Choice

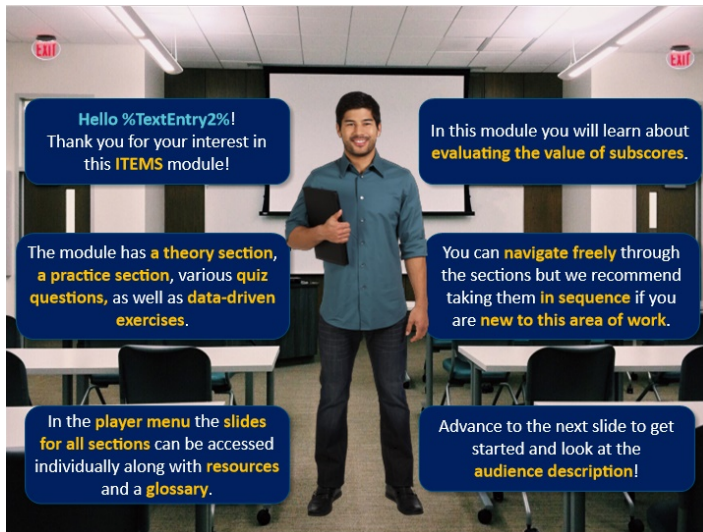


Module Introduction

Main Menu

Welcome %TextEntry2%!

1.6 Overview




1.7 Target Audience

Target Audience

Anyone who would like a **gentle statistical introduction** to this topic:

- graduate students and faculty in Master's, Ph.D., or certificate programs
- psychometricians and other measurement professionals
- data scientists / analysts
- research assistants / scientists
- technical project directors
- assessment development leads



However, we hope that you find the information in this module **useful no matter what your official title, role, or responsibility** in an organization is!

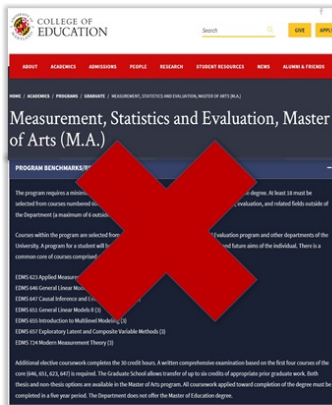
1.8 Expectations (I)



Let's discuss expectations....

1.9 Expectations (II)

ITEMS Modules in Context



1.10 Prerequisites

Prerequisites

To get the most out of this module it is beneficial to have the following **background knowledge and basic experiences**:

- Working knowledge of **basic statistical concepts** (e.g., random variables, distributions, and summary statistics)
- Working knowledge of **basic measurement concepts** from classical test theory (e.g., true score, reliability, standard error)
- Working knowledge of **basic principles of dimensionality analysis** (e.g., factor analyses, multidimensional item response theory)
- Basic experience with **running code in R**

However, the module author walks you through **basic ideas** of all key procedures to **support your learning**.

1.11 Resources

Resources

Sinharay, S., Puhan, G., and Haberman, S. (2011). An NCME Instructional Module on Subscores. *Educational Measurement: Issues and Practice*, 30 (3), 29 - 40. Available online at <https://ncme.elevate.commpartners.com/>

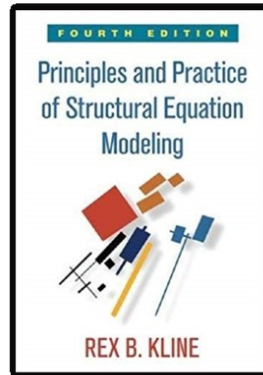
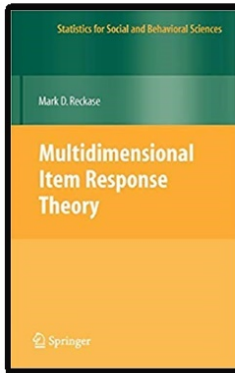
Module Citation



Additional References

References (Slide Layer)

Additional References



Back

1.12 Learning Objectives

Learning Objectives



1. Understand what subscores are and why they are of interest to users
2. Understand the need to assess the quality of the subscores
3. Know about the methods to assess the quality of subscores
4. Understand how one can decide whether to report subscores for a test

1.13 Main Menu

The image shows a 'Main Menu' interface. At the top, there is a dark blue header with a home icon and a tree icon on the left, and the text 'Main Menu' in yellow on the right. Below the header, the menu is divided into two main sections: 'Theory' and 'Practice'. The 'Theory' section is enclosed in a light green rounded rectangle and contains three items: '01 Conceptual Foundations [10 Minutes]', '02 Multivariate Methods [15 Minutes]', and '03 Haberman Method [30 Minutes]'. The 'Practice' section is enclosed in a light blue rounded rectangle and contains two items: '04 Derivativity [10 Minutes]' and '05 Quizzes [10 Minutes]'. A large red prohibition sign (a circle with a diagonal slash) is overlaid on the 'Practice' section. At the bottom of the menu area, there is a blue-bordered box containing the text: 'This slide deck contains only the slides from the first three content sections and the module introduction without any videos slides.'



2. Section 1: Conceptual Foundations


2.1 Cover: Section 1



Section 1:
Conceptual Foundations
[10 Minutes]

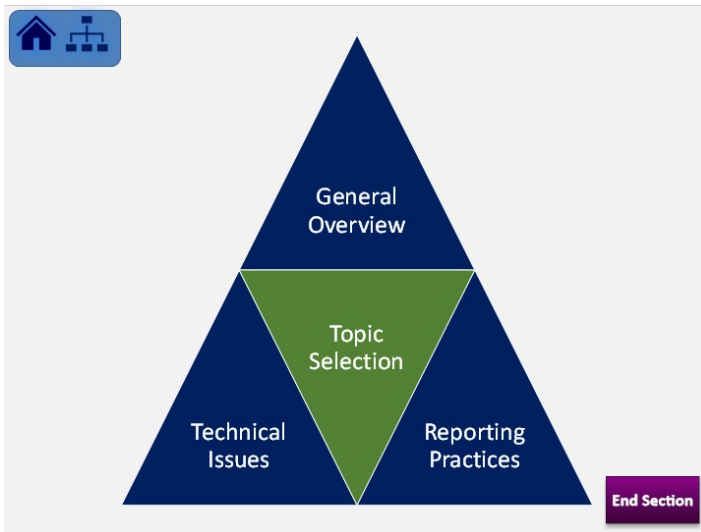
2.2 Objectives: Section 1

  **Learning Objectives**



1. Understand what subscores are
2. Understand why subscores are of interest to test users
3. Know about the current practice on reporting subscores
4. Understand the need to assess the quality of subscores

2.3 Topic Selection





2.4 Bookmark: General Overview






2.5 Bookmark: Reporting Practices



2.6 Definition of Subscores

  **What Are Subscores?**

- Scores on **a cluster of items** (i.e., subtest)
- Subtests usually correspond to different **content areas / skills**
- Reported for **large-scale assessments** (e.g., SAT®, ACT®, Praxis®,...)


  

1 2 3

Click on logos to view sample reports

Praxis sample (Slide Layer)

PRAXIS Score Report: Subscores


Test Taker
Score Report

Test / Test Category *	Your Raw Points Earned	Average Performance Range **
ELEMENTARY EDUCATION: CONTENT KNOWLEDGE (5018)		
I. ENGLISH LANGUAGE ARTS	35 out of 42	25-32
II. MATHEMATICS	32 out of 36	19-27
III. SOCIAL STUDIES	15 out of 20	9-13
IV. SCIENCE	16 out of 21	11-16
ELEMENTARY EDUCATION: CURRICULUM, INSTRUCTION, AND ASSESSMENT (5017)		
I. READING AND LANGUAGE ARTS	33 out of 37	23-29
II. MATHEMATICS	26 out of 31	19-25
III. SCIENCE	15 out of 20	11-15
IV. SOCIAL STUDIES	14 out of 17	9-13
IV. ART, MUSIC, AND PHYSICAL EDUCATION	13 out of 15	8-12

Back

SAT sample (Slide Layer)

SAT Score Report: Subscores



Aaron Farnsworth

Your Total Score

1510

400 to 1600

99th Nationally Representative Sample Percentile | 99th SAT User Percentile — National

Essay Scores

8 | 2 to 8

Reading

8 | 2 to 8

Analysis

8 | 2 to 8

Writing

Cross-Test Scores | 10 to 40

37

Analysis in History/Social Studies

36

Analysis in Science

Subscores | 1 to 35

15
Command of Evidence

13
Words in Context

15
Expression of Ideas

13
Standard English Conventions

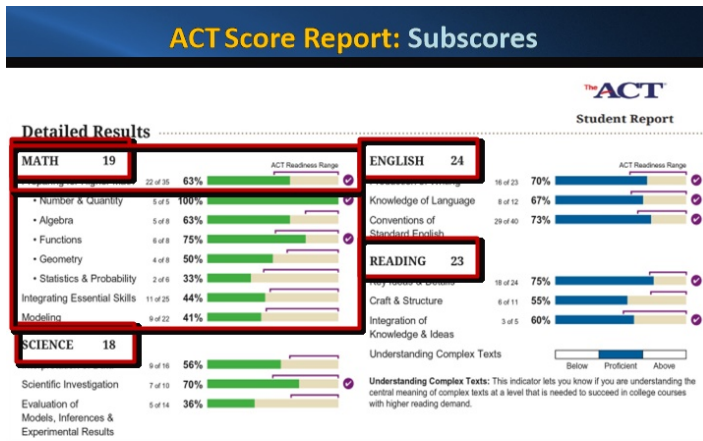
14
Heart of Algebra

15
Problem Solving and Data Analysis

15
Passport to Advanced Math

Back

ACT sample (Slide Layer)




Back

2.7 Benefits of Subscores (I)


Remedial Benefits:

Provide particular information about a learner's knowledge, skill, and abilities





Instructional Benefits:

Can be used by teachers and/or administrators to address students' specific academic needs



2.8 Benefits of Subscores (II)



Benefits of Subscores (II)

For learners:

Better plan independent studies to overcome weaknesses in skills, misconceptions, or strategies for low subscores



For teachers:

Better plan instructional efforts to help students overcome weaknesses for areas associated with low subscores

For states and other sponsors:

- Evaluate instructional program effectiveness
- Influence policies for training and educational programs


2.9 Legal Foundations



Legal Motivations

PUBLIC LAW 114-95—DEC. 10, 2015 129 STAT. 1827

“(x) produce individual student interpretive, descriptive, and diagnostic reports, consistent with clause (iii), regarding achievement on such assessments that allow parents, teachers, principals, and other school leaders to understand and address the specific academic needs of students, and that are provided to parents, teachers, and school leaders, as soon as is practicable after the assessment is given, in an understandable and uniform format, and to the extent practicable, in a language that parents can understand;



References

Ref (Slide Layer)

References

No Child Left Behind Act

The No Child Left Behind Act authorizes several federal education reauthorizations of the [Elementary and Secondary Education Act](#).

Under the 2002 law, states are required to test students in reading and math annually. The major focus of No Child Left Behind is to close student achievement gaps and provide significant opportunity to obtain a high-quality education. The U.S. Department of Education has identified several key features of the law:

- **Accountability:** To ensure these students who are disadvantaged and at risk of dropping out.
- **Flexibility:** Allows school districts flexibility in how they use federal funds.
- **Research-based education:** Emphasizes educational program research.
- **Parent options:** Increases the choices available to the parents.

NCLB requires each state to establish *state academic standards* and *accountability requirements* called *Adequate Yearly Progress (AYP)* from the U.S. Department of Education on August 8, 2008.

In its current iteration, NCLB formally expired on Sept. 30, 2007.

NCLB

Elementary and Secondary Education Act

Every Student Succeeds Act (ESSA) Implementation

For updated ESSA Consolidated Plan Drafting and Comment information

The Every Student Succeeds Act (ESSA) replaced No Child Left Behind (NCLB) as the federal law governing the process of implementing the law, beginning with the identification of schools for improvement. More information will be posted here as it becomes available.

2018 School District Accountability Annual Form
[The Every Student Succeeds Act \(ESSA\)](#)
[Washington's ESSA Consolidated Plan](#)

The Washington School Improvement Framework (WSIF)

The WSIF is the framework for accountability in Washington state. Each school in the Superintendent of Public Instruction (SPI) has identified schools for additional support.

- [Go to the Washington School Improvement Framework](#)
- [WSIF Handbook for all Districts | Overview](#)
- [WSIF Handbook | Information](#)
- [Frequently Asked Questions about the WSIF | District | District | District | District | District | District](#)

ESSA Implementation 101 - ESSA Implementation 101 is an overview of ESSA accountability system.

ESSA

Click on images to go to web sites



Back

2.10 Bookend: General Overview




This is the end of this topic.



2.11 Recommendations (I)


  **Recommendations (I)**

- ✓ Provide **easy-to-read narrative summaries** of subscores
- ✓ Provide **baseline information** to make interpretations meaningful
- ✓ Provide information on the **precision** of subscores
- ✓ Provide **clear instructions** on how subscores should be used



2.12 Reporting Practices (II)

  **Reporting Practices (II)**



Roberts & Gierl (2010) provided guidelines for reporting and presenting **diagnostic scores**.

The guidelines **apply to subscore reporting**, which are a type of diagnostic score.

[Reference](#)

Reference (Slide Layer)



EDUCATIONAL MEASUREMENT: ISSUES AND PRACTICE

NCME

Educational Measurement: Issues and Practice
Fall 2010, Vol. 29, No. 3, pp. 25–39

Developing Score Reports for Cognitive Diagnostic Assessments

Mary Roduta Roberts and Mark J. Gierl, *Centre for Research in Applied Measurement and Evaluation, University of Alberta*

This paper presents a framework to provide a structured approach for developing score reports for cognitive diagnostic assessments (CDAs). Guidelines for reporting and presenting diagnostic scores are based on a review of current educational test score reporting practices and literature from the area of information design. A sample diagnostic report is presented to illustrate application of the reporting framework in the context of one CDA procedure called the Attribute Hierarchy Method. Integration and application of interdisciplinary techniques from education, information design, and technology are required for effective score reporting. While the AHP is used in this paper, this framework is applicable to any attribute-based diagnostic testing method.

Keywords: cognitive diagnostic assessment, design principles, score reporting

Educational tests should provide meaningful information to guide student learning. The recent emphasis on understanding the psychology underlying test performance has led to developments in cognitive diagnostic assessment (CDA) (e.g., Leighton & Gierl, 2007a; Mislevy, 2006), which integrates cognitive psychology and educational measurement for the purposes of enhancing learning and instruction. A CDA is specifically designed to measure a student's knowledge structures and processing skills. In contrast with reporting a small number of content-based subareas, typical of most current educational test score reports, the results of a CDA yield a profile of scores with specific information about a student's testing process, there has been a paucity of research in this area. The available body of research on test score reporting has centered on large-scale reporting of aggregate-level results (i.e., at district, state, and national levels) for accountability purposes in the United States (Jager, 1998; Linn & Dunbar, 1992). Fewer studies have focused on student-level score reporting features (Goodman & Hambleton, 2004; Trout & Hyde, 2006). General conclusions drawn from these studies are not encouraging claiming that score reports are difficult to read and understand (Hambleton & Slater, 1997), often lead to inferences not supported by the information presented (Koretz & Robert, 1999) and



2.13 Recommendations (II)



A diagnostic report should include three sections:

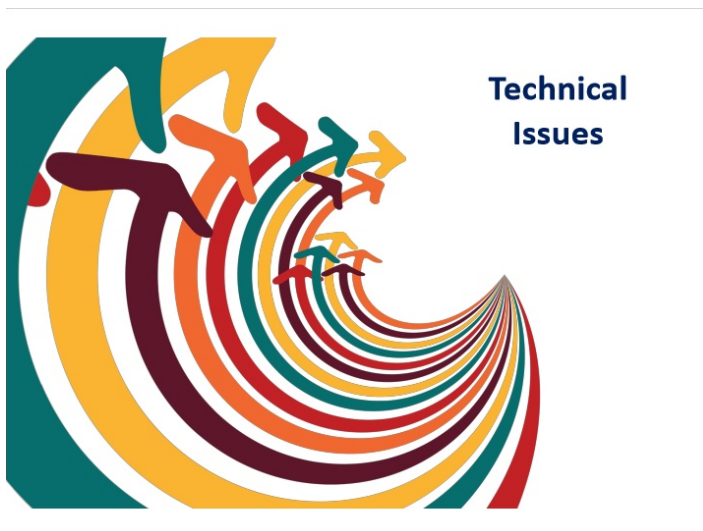
- **Top:** Overview of the content of the report
- **Middle:** Diagnostic information with item-level performance
- **Bottom:** Narrative summary across the subareas





2.14 Bookend: Reporting Practices






2.15 Bookmark: Technical Issues

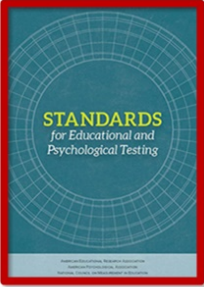


2.16 Quality Standards (I)



  **Quality Standards (I)**

- Subscores have to satisfy **professional standards** to be reportable
- Requires adequate **reliability, validity, and distinctness** of subscores



2.17 Reporting Practices (I)

  **Reporting Practices (I)**

Goodman & Hambleton (2004) did a **comprehensive review and critique** of score reporting practices from **large-scale assessments**

They found that subscores were mostly reported as **number-correct / raw scores, percent-correct scores, or percentile raw scores**

[Reference](#)

Reference (Slide Layer)



Reference

SPR-030-06-01-000001-01 EDUCATION, 7/22, 173, 229
Copyright © 2006, Lawrence Erlbaum Associates, Inc.

**Student Test Score Reports
and Interpretive Guides: Review
of Current Practices and Suggestions
for Future Research**

Dean P. Goodrum and Ronald K. Hambleton
*Center for Educational Assessment
University of Massachusetts Amherst*


A critical, yet often neglected, component of any large-scale assessment program is the reporting of test results. In the past decade, a body of evidence has been assembled that raises concerns over the ways in which these results are reported to and understood by their intended audiences. In this study, current approaches for reporting standardized results on large-scale assessments were investigated. Recent student test score reports and interpretive guides from 11 states, three U.S. commercial testing companies, and two Canadian provinces were reviewed. On the basis of test score reporting research, testing standards, and the requirements of the No Child Left Behind Act of 2001, a number of promising and potentially problematic features of these reports and guides are identified, and recommendations are offered to help enhance future score-reporting designs and to inform future research in this important area.

Large-scale assessments have played a prominent role for many years in America's kindergarten to Grade 12 school systems (Hamilton & Krentz, 2002; Linn, 1998), informing a wide range of national, state, and local reform efforts designed to improve student learning. Over this time, a great amount of attention has been directed toward the creation of technically sound assessments that can stand up to intense public and professional scrutiny. Considerably less attention, however, has been given to ways in which the results of the assessments are organized, reported,

Requests for reprints should be sent to Dean P. Goodrum, Center for Educational Assessment, 111 Shivers Center, University of Massachusetts, Amherst, MA 01003. E-mail: dgoodrum@psych.umass.edu

[Back](#)

2.18 Quality Standards (II)



Quality Standards



"When a test provides more than one score, the distinctiveness and reliability of the separate scores should be demonstrated."

**Standard 1.14 of the Standards for Educational and Psychological Testing
(AERA, APA, & NCME, 2014)**

"..the decision to provide subscores to candidates should be made carefully and information should be provided to facilitate proper interpretation".


**Chapter 11 (p. 176) of the Standards for Educational and Psychological Testing
(AERA, APA, & NCME, 2014)**

2.19 Technical Issues (I)

  **Technical Issues (I)**



Statistical Properties

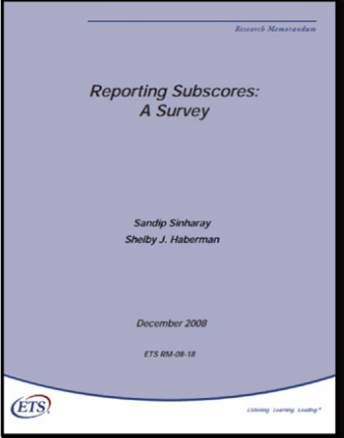
- ✓ Many tests include subscores based on a **few items**
- ✓ Few tests provide **precision / reliability** of reported subscores
- ✓ Reliabilities of subscores are **often small** for operational tests



[Reference](#)



Reference (Slide Layer)

  **Reference**






[Back](#)

2.20 Technical Issues (II)



  **Technical Issues (II)**


- Educational tests are often constructed to measure a **single construct**
 - ➔ subscores from such tests are **not expected to be very reliable**

- Subscores on educational tests most often refer to a **broad domain**
 - ➔ few items are **unlikely to provide accurate / precise information**

2.21 Bookend: Technical Issues



This is the end of this topic.

2.22 Summary



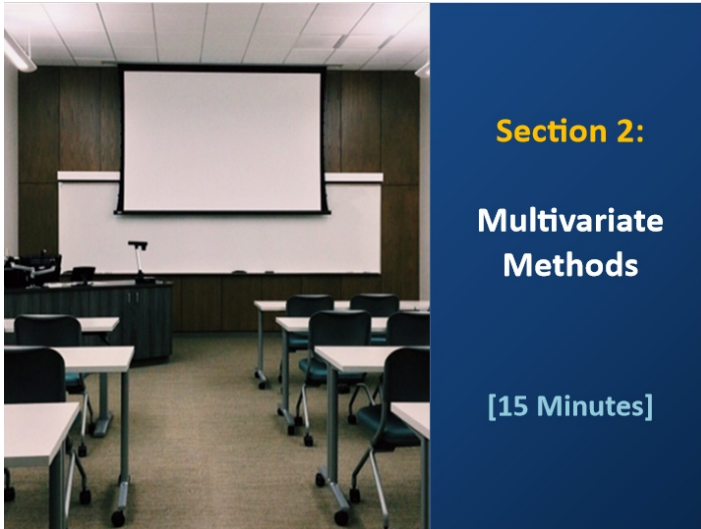
Summary

- **Everyone wants subscores** because of their potential remedial and instructional benefits
- **Subscores are reported** for many large-scale assessments and other learning environments
- It is important to report subscores only when they **satisfy professional quality standards**






3. Section 2: Multivariate Methods

3.1 Cover: Section 2





3.2 Objectives: Section 2

  **Learning Objectives**




1. Learn about the various methods to assess the quality of subscores
2. Understand how these methods work for a data set
3. Understand the differences of these methods



3.3 Introduction (I)

  **Methods Overview**

1. Descriptive statistics
2. Factor analysis
3. Multidimensional item response theory
4. Dimensionality-detection statistics




3.4 Introduction (II)

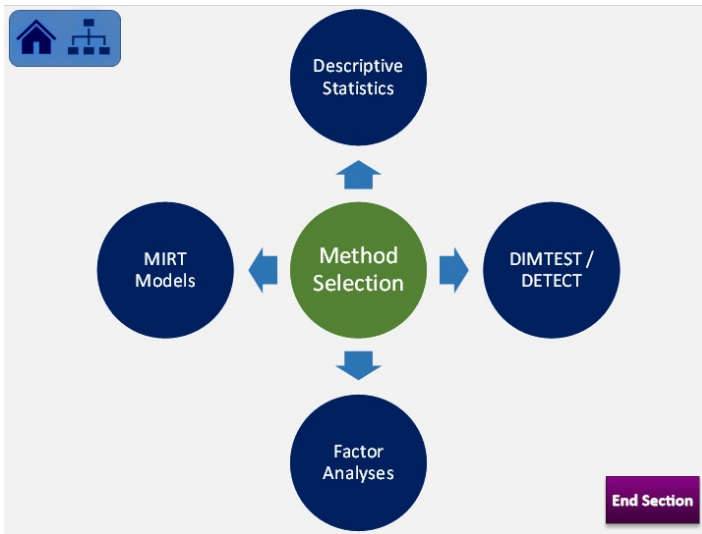
  **Methods to Compute Added Value**

Example data from an achievement test at the K-12 level

- **60** dichotomous items and about **4,000** examinees
- **3** content areas (**20** items each): Mathematics, Reading, Social Studies
- Subscores used to **identify strengths and weaknesses** of students
- Methods applied for assessing the **added value of subscores**





3.5 Methods Selection



3.6 Bookmark: Descriptive Statistics


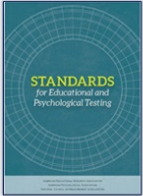


3.7 Principles (I)



  **Basic Principles**

Reporting of subscores is not justified

- × they are **highly correlated**
- × they do **not** have **adequate reliability**


3.8 Principles (II)

  **Implementation**

Do not report subscores if:

- × **pair-wise disattenuated correlations are large** (e.g., greater than 0.90)
- × **individual subscore reliabilities are low** (e.g., smaller than .80)

Not a rigorous test, but provides a quick/rough idea

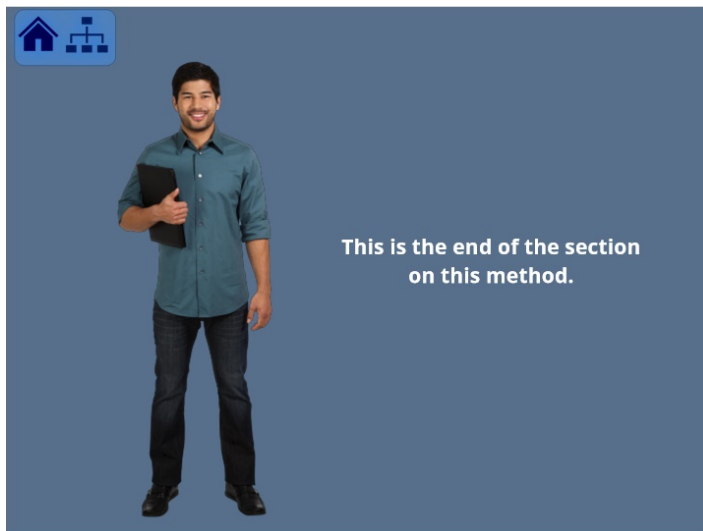


3.9 Example

	Math	Reading	Social Studies
Mathematics	0.81	0.64	0.64
Reading	0.78	0.83	0.68
Social Studies	0.79	0.83	0.80

Diagonal: Reliabilities
Upper off-diagonal: Simple correlations
Lower off-diagonal: Disattenuated correlations



3.10 Bookend: Descriptive Statistics




3.11 Bookmark: DIMTEST & DETECT



3.12 Methods

  **Methods**

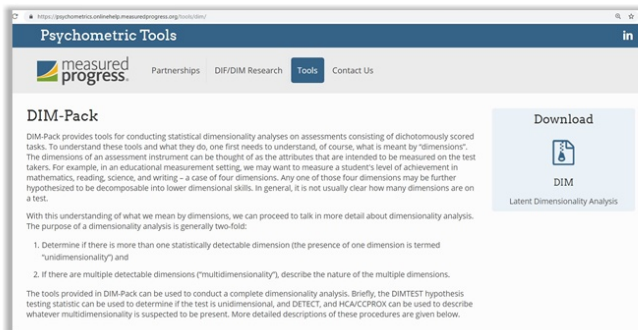
- Several (nonparametric) **dimensionality detection methods** exist
- This module section covers **two key methods**:
 - ✓ **DETECT**: Effect size (descriptive)
 - ✓ **DIMTEST**: Hypothesis test (inferential)

 [Reference](#)

Reference (Slide Layer)






3.13 Software




<https://psychometrics.onlinehelp.measuredprogress.org/tools/dim/>

3.14 Principles [DETECT] (I)



  **Basic Principle: DETECT**

 Items that measure the **same** dimension exhibit **positive** conditional covariances

Items that measure **different** dimensions exhibit **negative** conditional covariances 

[Reference](#)

Reference (Slide Layer)

  **Reference**

Journal of Educational Measurement
Fall 2006, Vol. 43, No. 3, pp. 215-243



Formulation of the DETECT Population Parameter and Evaluation of DETECT Estimator Bias

Louis A. Ronsson
Measured Progress
Orlem Yesin Ozbek
Gaziosmanpaşa University

The development of the DETECT procedure marked an important advancement in nonparametric dimensionality analysis. DETECT is the first nonparametric technique to estimate the number of dimensions in a data set, estimate an effect size for multidimensionality, and identify which dimension is predominantly measured by each item. The efficacy of DETECT critically depends on accurate, minimally biased estimation of the expected conditional covariances of all the item pairs. However, the amount of bias in the DETECT estimator has been studied only in a few simulated unidimensional data sets. This is because the value of the DETECT population parameter is known to be zero for this case and has been unknown for cases when multidimensionality is present. In this article, integral formulas for the DETECT population parameter are derived for the most commonly used parametric multidimensional item response theory model, the Beckas and McKinley model. These formulas are then used to evaluate the bias in DETECT by positing a multidimensional model, simulating data from the model using a very large sample size (to eliminate random error), calculating the large-sample DETECT statistic, and finally calculating the DETECT population parameter to compare with the large-sample statistic. A wide variety of two- and three-dimensional models, including both simple structure and approximate simple structure, were investigated. The results indicated that DETECT does exhibit statistical bias in the large-sample estimation of the item-pair conditional covariances, but, for the simulated tests that had 20 or more items, the bias was small enough to result in the large-sample DETECT almost always correctly partitioning the items and the DETECT effect size estimator exhibiting negligible bias.

[Back](#)

3.15 Principles [DETECT] (II)



  **Implementation: DETECT**

DETECT statistic (D)
Weighted mean of the conditional covariances among the item pairs

Magnitude of the DETECT statistic indicates the dimensionality



- $0 < D < .10$ = likely unidimensional (no subscores)
- $.10 \leq D < .50$ = weakly to moderately multidimensional
- $.50 \leq D \leq 1.00$ = strongly multidimensional

3.16 Example [DETECT]

  **Example: DETECT**



DETECT for the Achievement Test Data

DETECT = 0.42

- Data **moderately multidimensional**
- Subscores may be **reportable**

3.17 Principles [DIMTEST] (I)

  **Basic Principle: DIMTEST**

- Evaluate **data likelihood** (joint probability)
 - IF: likelihood that the item scores originated from an essentially unidimensional item pool is **high**
 - THEN: subscore reporting is **not justified**
- Compute **test statistic (T)** and conduct **statistical inference** (hypothesis test, confidence interval)

[Reference](#)

Reference (Slide Layer)

  **Reference**

 **JEM** JOURNAL OF EDUCATIONAL MEASUREMENT  **NCME** NATIONAL COUNCIL ON MEASUREMENT FOR EDUCATION
Journal of Educational Measurement
Winter 2010, Vol. 47, No. 4, pp. 413-431



Formulation of a DIMTEST Effect Size Measure (DESM) and Evaluation of the DESM Estimator Bias

Minhee Seo
University of North Carolina at Greensboro
Louis A. Roussos
Measured Progress

DIMTEST is a widely used and studied method for testing the hypothesis of test unidimensionality as represented by local item independence. However, DIMTEST does not report the amount of multidimensionality that exists in data when rejecting its null. To provide more information regarding the degree to which data depart from unidimensionality, a DIMTEST-based Effect Size Measure (DESM) was formulated. In addition to detailing the development of the DESM estimate, the current study describes the theoretical formulation of a DESM parameter. To evaluate the efficacy of the DESM estimator according to test length, sample size, and correlations between dimensions, Monte Carlo simulations were conducted. The results of the simulation study indicated that the DESM estimator converged to its parameter as test length increased, and, as desired, its expected value did not increase with sample size (unlike the DIMTEST statistic in the case of multidimensionality). Also as desired, the standard error of DESM decreased as sample size increased.


[Back](#)

3.18 Principles [DIMTEST] (II)



  **Implementation: DIMTEST**

Compute the DIMTEST (T) statistic and p -value for hypothesis test

H_0 : data are **essentially unidimensional**
 H_a : data have **more than one** underlying dimension




3.19 Principles [DIMTEST] (III)

  **Implementation: DIMTEST**



Interpret the outcome of the test:

IF: The observed T statistic is **larger** than a chosen cutoff (e.g., 1.64 for a 5% type-I error rate) **reject H_0**

THEN: The data are **likely multidimensional** and the subscores are likely to have **added value**





3.20 Example [DIMTEST]

  **Example: DIMTEST**



DIMTEST for the Achievement Test Data
(exploratory mode)


$T = 4.84, p\text{-value} < .05$

- Data are **likely multidimensional**
- Subscores may be **reportable**

3.21 Bookend: Descriptive Statistics





This is the end of the section
on this method.

3.22 Bookmark: Factor Analysis



3.23 Principles (I)

  **Basic Principles**

IF: the scores on the test items can be mostly explained by one latent variable
THEN: the test primarily measures one overall ability

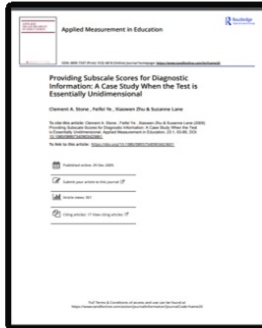
IF: it takes a few latent variables to explain the scores on the test items
THEN: the test measures multiple abilities

Each latent variable is a **statistical dimension/factor** and could correspond to a **subscore for reporting**

[References](#)

References (Slide Layer)

References



Stone et al. (2010)



Sinharay et al. (2007)

Back

3.24 Principles (II)

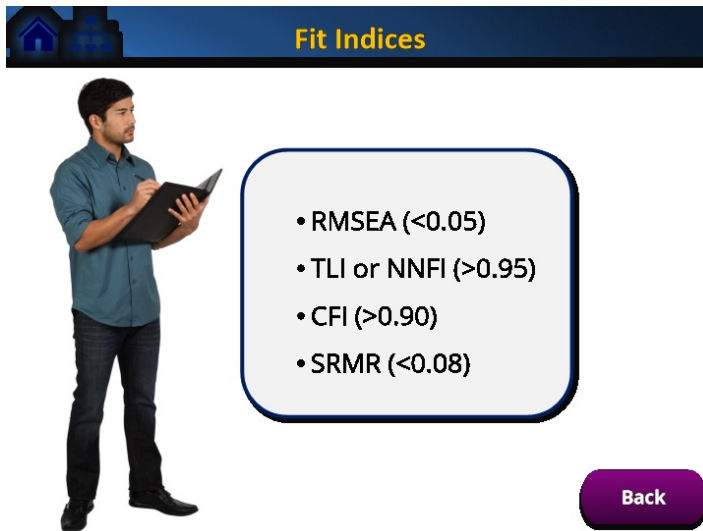
Implementation (I)

- Fit **factor analysis** models to the data (**confirmatory factor analysis** preferred, **exploratory factor analysis** possible)
- Evaluate **fit indices** for all models (**1-factor and above**) to obtain **statistical evidence** regarding the number of factors
- The **R package psych** includes several fit indices for CFA but one can use **other programs** such as AMOS, LISREL, and Mplus



Fit Indices

Fit Indices (Slide Layer)

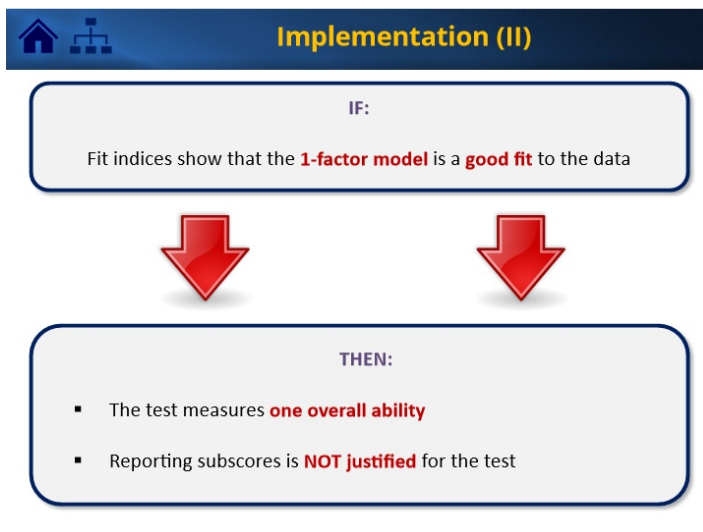


Fit Indices

- RMSEA (<0.05)
- TLI or NNFI (>0.95)
- CFI (>0.90)
- SRMR (<0.08)

Back

3.25 Principles (III)



Implementation (II)



IF:

Fit indices show that the **1-factor model** is a **good fit** to the data

THEN:



- The test measures **one overall ability**
- Reporting subscores is **NOT justified** for the test

3.26 Principles (III)

  **Implementation (III)**

IF:



A model with **multiple factors** is a better fit than a **1-factor model**

THEN:


Reporting **several subscores** is probably **justified**

3.27 Example

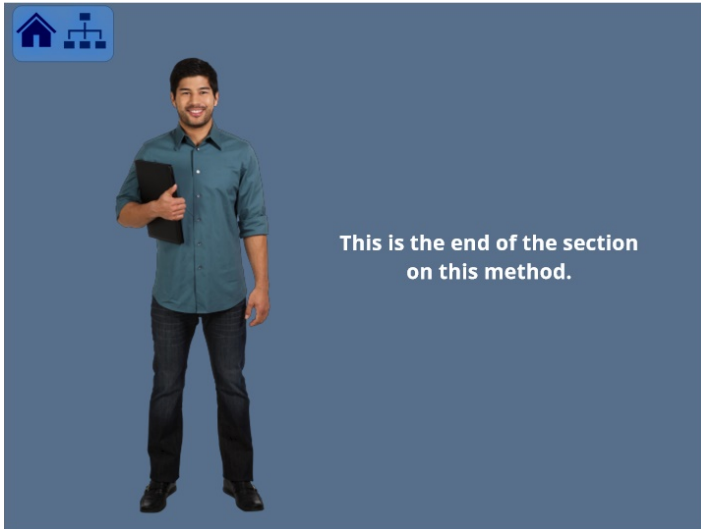
  **Example**

Fit indices for the Achievement Test Data
(lower RMSEA = less error = better fit)

RMSEA for 1-factor model:	0.06
RMSEA for 3-factor model:	0.04





3.28 Bookend: Descriptive Statistics



3.29 Bookmark: MIRT




3.30 Principles (I)

  **Basic Principle**

Determine best-fitting model



of dimensions = 1 **NOT justified** to report subscores
of dimensions > 1 **Justified** to report subscores

Very similar in purpose to factor-analytic methods



[Reference](#)

Reference (Slide Layer)

  **References**

An NCFE Instructional Module on

Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests

Terry A. Ackerman, University of North Carolina-Greensboro
Mark J. Gierl, University of Alberta
Cindy M. Walker, University of Wisconsin-Milwaukee

Using educational and psychological tests are inherently multidimensional, meaning that they measure two or more dimensions or constructs. The purpose of this module is to describe how test practitioners and researchers can apply multidimensional item response theory (MIRT) to understand better what their tests are measuring. Also, describing the different components of ability are being assessed, and how this information can be applied back into the test development process. Procedures for conducting MIRT analyses from obtaining evidence that the test is multidimensional, to modeling the test as multidimensional, to illustrating the properties of multidimensional item response theory described from both a theoretical and a substantive level. This module also illustrates these procedures using data from a middle-grade mathematics achievement test. It concludes with a discussion of future directions in MIRT research.



Keywords: item response theory, multidimensional item response theory, test development and analysis

The purpose of this module is to describe how test practitioners and researchers can apply multidimensional item response theory (MIRT) to understand better what their tests are measuring. Also, describing the different components of ability are being assessed, and how this information can be applied back into the test development process. MIRT is used to model multidimensional tests, to illustrate how these tests are being measured, and to describe the relationship between the test and the construct being measured. The module also illustrates these procedures using data from a middle-grade mathematics achievement test. It concludes with a discussion of future directions in MIRT research.


Keywords: item response theory, multidimensional item response theory, test development and analysis

Back



3.31 Principles (II)

  **Implementation**

- Fit **unidimensional** and **multidimensional** IRT models to data
- Evaluate **relative fit** using **information indices** such as:
 - ✓ Akaike's information criterion (AIC)
 - ✓ Bayesian information criterion (BIC)





3.32 Principles (III)

  **Implementation**

IF:

Fit indices show that a **multidimensional model** fits the data better

THEN:

- The test measures **more than one overall ability**
- Reporting subscores is **likely justified** for the test

Ref (Slide Layer)

References

An NCMIE Instructional Module on
Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests
 Terry A. Ackerman, University of North Carolina-Greensboro
 Mark J. Gierl, University of Alberta
 Cindy M. Walker, University of Wisconsin-Milwaukee

Many educational and psychological tests are inherently multidimensional, meaning that they measure tests on several dimensions or constructs. The purpose of this module is to illustrate how test practitioners and researchers can apply multidimensional item response theory (MIRT) to understand better what their tests are measuring, how accurately the different components of ability are being assessed, and how this information can be applied back into the test development process. Procedures for understanding MIRT and how to use it to evaluate tests are described from both a theoretical and a substantive basis. This module also discusses these procedures using data from a sixth-grade mathematics achievement test. It concludes with a discussion of future directions in MIRT research.

Keywords: dimensionality, multidimensional item response theory, test development and analysis

The purpose of this article is to illustrate how test practitioners can use multidimensional item response theory (MIRT) to understand better what their tests are measuring, how accurately the different components of ability are being assessed, and how this information can be applied back into the test development process. Procedures for understanding MIRT and how to use it to evaluate tests are described from both a theoretical and a substantive basis. This module also discusses these procedures using data from a sixth-grade mathematics achievement test. It concludes with a discussion of future directions in MIRT research.

Keywords: dimensionality, multidimensional item response theory, test development and analysis

Many educational and psychological tests are inherently multidimensional, meaning that they measure tests on several dimensions or constructs. The purpose of this module is to illustrate how test practitioners and researchers can apply multidimensional item response theory (MIRT) to understand better what their tests are measuring, how accurately the different components of ability are being assessed, and how this information can be applied back into the test development process. Procedures for understanding MIRT and how to use it to evaluate tests are described from both a theoretical and a substantive basis. This module also discusses these procedures using data from a sixth-grade mathematics achievement test. It concludes with a discussion of future directions in MIRT research.

Keywords: dimensionality, multidimensional item response theory, test development and analysis

[Back to Main Slide](#)

3.33 Example

Example

Fit indices for the Achievement Test Data
 (lower fit index = less error = better fit)


Model	AIC	BIC
Unidimensional	239,136	239,890
Multidimensional	237,255	238,027



3.34 Bookend: Descriptive Statistics



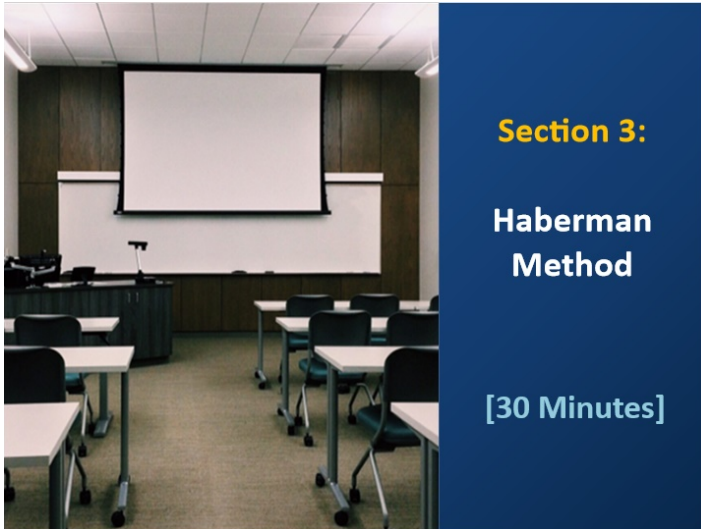
3.35 Summary

 **Summary**



- Various **multivariate methods** are available for determining the added value of subscores through **dimensionality evaluation**
- The methods rely on various information ranging from **simple summary statistics** to sophisticated **statistical models**
- The various methods lead to **similar conclusions** for the **example data** set, but that **may not always** be the case
- In cases of **disagreement**, more **detailed analyses** of **loading patterns** and input from **subject-matter experts** is needed


4. Section 3: Haberman's Method

4.1 Cover: Section 3



4.2 Objectives: Section 3

  **Learning Objectives**





1. Understand the basics of the Haberman method to determine the added value of subscores




2. Understand how to apply the Haberman method to a data set

3. Understand how to make a decision on the added value of subscores using the Haberman method



4.3 Principles (I)



Basic Principles (I)



- Each item contributes to only one subscore
 **Simple loading structure**
- All students were administered all items on the test
 **Fully crossed design**
- Key assumptions of classical test theory likely hold for the data
 **No complex response patterns**

4.4 Principles (II)



Basic Principles (II)

- Based on principles from **classical test theory**
- Based on **simple data summaries** such as **reliabilities** and **covariances** of the subscores
- Provides a **clear decision rule** whose logic is transparent
- Implemented in the **subscore package in R**



References (Slide Layer)



Educational Measurement: Issues and Practice
Fall 2013, Vol. 36, No. 3, pp. 48-50

An NCTE Instructional Module on Subscores

Stanley S. Shinar, Gautam Puhani, and Shelby J. Huberman, Educational Testing Service

The purpose of this (ETPS) module is to provide an introduction to subscores. First, examples of subscores from an operational test are presented. Then, a review of methods that can be used to examine if subscores have adequate psychometric properties is presented. It is demonstrated, using results from operational and simulation data, that subscores based on the analysis of a sufficient number of items and have to be sufficiently shared from each other to have adequate psychometric quality. It is also demonstrated that several operationally reported subscores do not have adequate psychometric quality. Recommendations are made for those interested in reporting subscores for educational tests.

Keywords: augmented subscore, classical test theory, diagnostic scores, item response theory, mean squared error

What Are Subscores and Why Should Anyone Care About Them?

Figure 1 shows a hypothetical score report of an operational test (Mary D. Pappas) in the Educational Testing Service (ETS) database that is designed for prospective teachers. The test is designed to measure the knowledge and skills of candidates for entry into the teaching profession. The test is composed of 100 items, and the total score is 100. The test is designed to measure the knowledge and skills of candidates for entry into the teaching profession. The test is composed of 100 items, and the total score is 100. The test is designed to measure the knowledge and skills of candidates for entry into the teaching profession. The test is composed of 100 items, and the total score is 100.

Copyright © 2013 by the National Council on Measurement in Education 49





4.5 Principles (III)



Observed score	True score
Total Score	True Total Score
Math Subscore	True Math Subscore
Reading Subscore	True Reading Subscore



A subscore has **added value** only if the corresponding true subscore can be **predicted better** by the **subscore** than by the **total score**

4.6 Principles (IV)



  **Basic Principles (IV)**

Original Form	Parallel Form
Total Score	Total Score
Math Subscore	Math Subscore
Reading Subscore	Reading Subscore

A subscore has **added value** only if it can **better predict** the **corresponding subscore** on a **parallel form** than does the **total score**

 Can be evaluated through single form 

4.7 Statistics


  **Computation**

For each subscore, calculate **two quantities**:

Reliability: Squared correlation between subscore and true subscore

PRMSE: Squared correlation between total score and true subscore

A subscore has added value if **Reliability > PRMSE**



Formulas

Formulas (Slide Layer)

Formulas

$$\text{PRMSE}_{total} = \rho^2(s_t, x) = \text{[redacted]} \times \text{Total score reliability}$$
$$\rho^2(s_t, x) = \frac{\text{Cov}(s_t, x)^2}{V(s_t)V(x)}$$

Notation **Back**



Notation (Slide Layer)

Notation

Notation	Quantity
PRMSE_{total}	PRMSE for the total score
$V(z)$	Variance of the variable z
$\text{Cov}(y, z)$	Covariance between variables y and z
$\rho^2(y, z)$	Squared correlation between variables y and z
s	Observed subscore
s_t	True subscore
x	Observed total score
x_t	True total score



Formulas **Back**

4.8 Example

  **Example**


Subscore	Reliability Subscore	PRMSE Total Score
Mathematics	0.81	0.77
Reading	0.83	0.81
Social Studies	0.80	0.81

4.9 Survey Results (I)

  **Survey of Applied Practice**

Application of Haberman method to data from 25 large-scale test:

- **16 tests** had no subscores with added value
- **9 tests** had subscores with added value:
 - based on **at least 20 items**
 - dimensions were **sufficiently distinct**



[Reference](#)

References (Slide Layer)

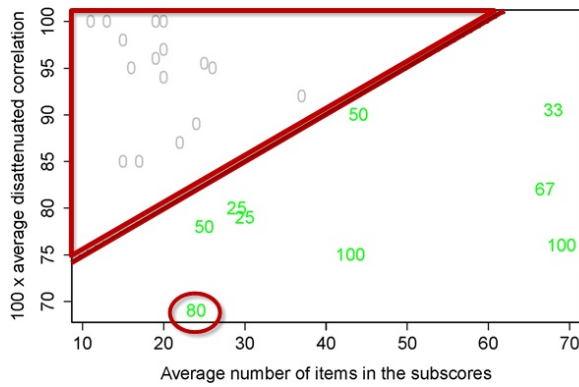
Reference



Back

4.10 Survey Results (II)

Graphical Illustration (I)



References (Slide Layer)

References

JEM JOURNAL OF EDUCATIONAL MEASUREMENT
NCME
Journal of Educational Measurement
 Summer 2010, Vol. 47, No. 2, pp. 130-134

How Often Do Subscores Have Added Value? Results from Operational and Simulated Data

Sandip Sinharay
 Educational Testing Service

Recently, there has been an increasing level of interest in subscores for their potential diagnostic value. Haberman suggested a method based on classical test theory to determine whether subscores have added value over total scores. In this article I first provide a rich collection of results regarding when subscores were found to have added value for several operational data sets. Following that I provide results from a detailed simulation study that examines what properties subscores should possess in order to have added value. The results indicate that subscores have to satisfy strict standards of reliability and correlation to have added value. A weighted average of the subscores and the total score was found to have added value more often.

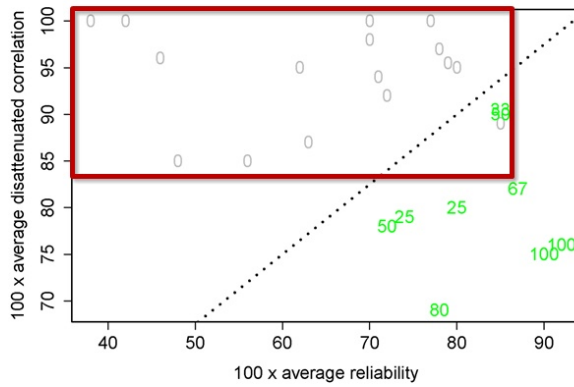
There is an increasing interest in subscores, which are scores on subtests, because of their potential diagnostic value. Individual examinees want to know their strengths and weaknesses in different content areas to plan for future remedial work. States and academic institutions such as colleges and universities often want a summary of performance for their students to better evaluate their training and focus on areas that need instructional improvement (Haladyna & Kramer, 2004). The U.S. Government's No Child Left Behind (NCLB) Act of 2001 demands, among other things, that students should receive diagnostic reports that allow teachers to address their specific academic needs; subscores could be used in such a diagnostic report. The total score, consisting of a single number, is often argued to be too deterministic, showing how the examinee's abilities vary over the different subtests may be of more use. Finally, it may be possible to improve predictive validity by using the subscores.

Despite this apparent usefulness of subscores, certain important factors must be considered before making a decision on whether to report subscores at either the individual or institutional level. According to Standard 5.17 of the Standards for

Back to Main Slide

4.11 Survey Results (III)

Graphical Illustration (II)



References (Slide Layer)

References



Back to
Main Slide

4.12 Extensions

Extensions & Further Studies

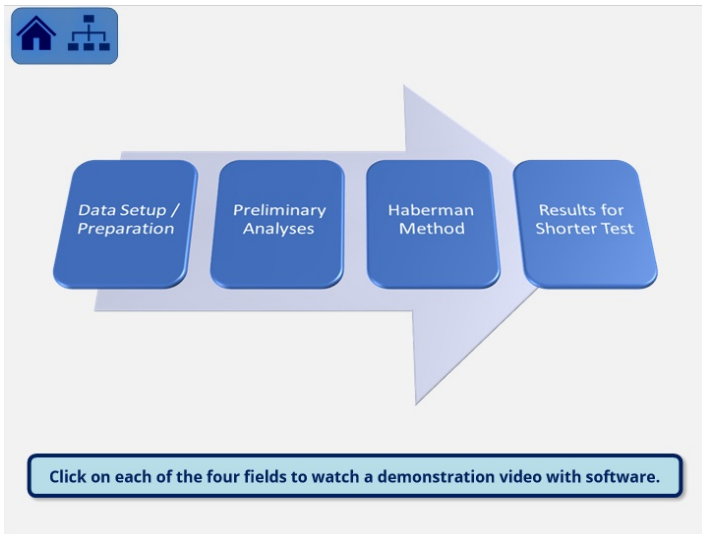
- Augmentation of subscores (Haberman, 2008)
- Aggregate-level subscores (Haberman et al., 2009)
- Subscores based on MIRT models (Haberman & Sinharay, 2010)
- Subgroups and fairness (Haberman & Sinharay, 2014)

Click on the individual titles above to
get to the journals for these articles

(paid access)



4.13 Video Illustrations



4.14 Module Cover (END)

