Selecting and Designing Instruments: Item Development, Reliability, and Validity

John D. Hathcoat, PhD & Courtney B. Sanders, MS, Nikole Gregg, BA

This document consists of an introduction to the selection and design of instruments in educational assessment. An overview of basic considerations in selecting/designing an instrument, item development, reliability, and validity is provided.

Contents

INTRODUCTION

Educational assessment may be defined as the systemic collection of information to make evidence-based decisions about the quality and/or effectiveness of educational programs. These programs are designed to enhance student learning and/or development. Two mistakes may be made with respect to evidence-based decisions: (1) concluding that an ineffective program is actually effective and (2) concluding that an effective program is actually ineffective. Various strategies may be used to minimize these problems; however, this guide focuses on the selection and design of an assessment instrument. Issues with respect to the selection and design of an assessment instrument arguably exacerbate both mistakes.

It is difficult, if not entirely impossible, to appropriately select and/or design an assessment instrument without a basic understanding of reliability and validity (i.e. measurement theory). Many assessment practitioners enter into this field with limited training in measurement theory. Even within various disciplines in education and the social sciences, training in measurement theory is limited at best. This issue is further problematized by common misconceptions about reliability and validity found within peer-reviewed literature.

Recognizing the abovementioned issues served as a motivation to write a brief handbook about the concepts of reliability and validity with respect to instrument selection/design. *This handbook was written while working from the assumption that readers would have a beginning level of statistics/methodology knowledge.* For example, we assume that the reader has a basic understanding of some statistical concepts, such as a correlation.

The terms "instrument" and "test" are used interchangeably throughout the document. These terms are used in a broad sense to indicate any systematic or standard procedure used to obtain indications of a skill, process, or attribute of interest. As we use the terms, instruments/tests refer to both "authentic" assessments, such samples of student coursework, and multiple choice exams.

The contents of the handbook are outlined as follows: (a) considerations in selecting versus designing an instrument, (b) item development, (c) reliability, and (d) validity. The handbook concludes by providing general recommendations aiming to synthesize the presented information.

Various decisions had to be made about what to include/exclude. Entire books are written about some of these topics; thus we sought to provide a conceptual introduction to each topic in sufficient detail for a reader to make informed decisions. *This handbook should be viewed as a supplement to more technical texts.* Additional resources about selected topics are provided for interested readers at the end of the document.

ACKNOWLEDGMENTS

We would like to thank all faculty members and graduate assistants in the Center for Assessment and Research Studies (CARS) at James Madison University. Our gratitude extends to both past and present members of CARS. This document, in many respects, has been influenced by years of prior work conducted by CARS. For example, faculty members and students have gathered numerous documents to serve as resources during assessment consultation. When writing this handbook, we had the luxury of drawing upon a wealth of existent resources and years of experience gained by a community of experts within this field. We would like to specifically thank Dr. Sara Finney for creating resources which we drew upon when writing the section about common mistakes when writing items.

SELECTING AND/OR DESIGNING AN ASSESSMENT INSTRUMENT

As an assessment practitioner, one is often faced with the challenge of deciding whether to select an existing instrument or to develop one's own instrument. Practitioners are often tempted to design their own instruments without first considering whether an existing instrument is available. There are a few issues to consider when making such a decision.

Perhaps most importantly, practitioners should not spend time "reinventing the wheel." A review of literature should always occur prior to selecting/designing an instrument. If an adequate and appropriate instrument exists then it is a waste of resources to develop a new instrument.

Direct versus indirect measures, verb-instrument agreement, as well as outcome-instrument maps can be used to *initially* examine the adequacy and appropriateness of an instrument. Each consideration is addressed in turn. Reliability and validity are addressed later in the handbook.

Direct versus Indirect

The terms "direct" and "indirect" are frequently used in the literature to describe an instrument. Unfortunately these terms are somewhat misleading and may result in stakeholders having undue confidence in assessment results. *There is no such thing as a direct measure of student learning. Direct measurement, strictly speaking, is a misnomer.*

The "directness" of an instrument is better conceived along a continuum ranging from those that are, relatively speaking, more or less direct. For example, students may be asked about their critical thinking skills, which is less direct assessment than an assignment asking them to critique a journal article. If one is interested however, in student beliefs about their critical thinking skills then the self-report measure is considered more direct than the assignment critique. *Measures that are more direct are always preferable to those that are less direct.*

Verb-Instrument Agreement

Student learning outcomes (SLO), or what is often referred to as objectives, includes an action verb indicating what a student is expected to know, think, or do as a result of program participation. Each verb acts as a hint about what type of instrument is appropriate. Three SLO's are provided below with a description of appropriate assessment strategies.

Bloom's taxonomy (Krathwohl, 2002) may be used when thinking about action verbs incorporated within an SLO. The verb in each SLO informs what type of response (closed versus open-ended) is needed to adequately assess the outcome. For example, the verb "create" in the Table below implies some form a performance assessment (e.g., constructed response such as a paper) is appropriate, which will likely be scored using a checklist or rubric.

When examining an assessment instrument one must check for verb-instrument agreement. In cases of disagreement, one may either modify the SLO or select/design a different instrument. Keep in mind that modifying an SLO simply to fit an instrument will likely require you to make curricular changes to the program (curriculum should be intentionally created to give students an opportunity to learn a stated outcome).

Student Learning Outcomes and Assessment Techniques

Student Learning Outcome	**Appropriate Assessment**	**Inappropriate Assessment**
Students will *create* a plan to contribute to the campus community.	Performance assessment (e.g. paper); scored as complete/incomplete or examined for quality using a rubric.	Multiple choice, Likert-type scales, etc.
Students will *list* three ways they will contribute to the campus community.	Open-ended question; could be scored as correct/incorrect or partial credit.	Multiple Choice, Likert-type scales, etc.
Students will *report* an increase in the sense of belonging to the campus community.	Likert-type scale using at least a pretest and posttest. Must find or create appropriate scale.	Multiple choice, performance assessment, etc.

Mapping to Each Learning Outcome

An instrument-learning outcome map is provided below for a single student learning outcome specified by a hypothetical program. An assessment instrument should align to each student learning outcome. A single instrument may map back to multiple outcomes or different instruments may be needed for each outcome. It is also important to document existing reliability and validity evidence (described in later sections).

Hypothetical Instrument-Outcome Map

Objective	***Instrument***	***Number of Items***	***Scoring***	***Reliability***	***Validity Evidence***
As a result of participating in X program,	Academic Requirements	7	All items are multiple	Internal consistency was estimated	Prior research has indicated

students will increase in the number of correctly identified graduation requirements at JMU.	Knowledge Scale		choice. A total score is calculated by summing the number of correct responses.	at approximately .80 among samples of similar students.	that there is a relationship between total scores and graduation (see Author 1 & Author 2, 2015).

(Include actual items in an Appendix).

Considerations when Selecting an Existing Instrument

You should be skeptical of the name given to an instrument by other researchers. These names can be misleading. In other words, just because someone labels an instrument a “critical thinking test” does not imply that it is a good measure of critical thinking.

You should always read the actual items before adopting an instrument. Students respond to items, not the name of a test. These items should reflect the objectives and curriculum designed to meet each objective. For example, if you are teaching critical thinking skills and the items primarily pertain to deductive logic then you should be teaching deductive logic if you adopt the test. If not, then you should a select different instrument. Other considerations are summarized below.

1. Consider the length of an instrument – this should be long enough to cover the breadth of what you want to measure without inducing fatigue.
2. Consider the potential cost of using an instrument. Some instruments are under propriety and cannot be used without special permission.
3. Evaluate reliability and validity evidence before using an existing instrument to assess your program. If you are using an instrument in a new way then you need to collect evidence to justify your proposed use of a test.

Considerations when Designing an Instrument

1. We strongly recommend for you to consult with an individual trained in measurement theory prior to creating your own instrument. However, there are circumstances in which this is simply unfeasible. In such cases, it is recommended to locate resources beyond this manual (see end of document).
2. Expect to spend about a year in the instrument development process. Many instruments take longer than a year to develop. Thus instrument development requires one to commit resources and time.

3. You will need to collect validity evidence to support your proposed interpretation of the scores for an intended use of a test. For example, if a test has been used in research to study externalizing problems in children it would need additional evidence to support its use to make placement decisions in school. This evidence can be time consuming to collect and may require you to make multiple modifications to the items before the instrument is ready for "official" administration.

Advantages and Disadvantages of Selecting versus Designing an Instrument

The table below provides a snapshot of some of the advantages and disadvantages of selecting versus designing an instrument.

	Advantages	**Disadvantages**
Existing Instrument	• Convenience • Instrument development is complete. • Existent reliability and validity evidence. • Comparisons to prior research and other groups are more feasible.	• Less than ideal alignment to outcomes. • Inadequate reliability and validity evidence. • Cost
New Instrument	• Better match to outcomes.	• Resource intensive (time consuming) • New validity evidence • Home-grown instruments limit external comparisons.

ITEM WRITING

Cognitive Items

Cognitive outcomes include particular forms of reasoning, knowledge, or processes students engage in to solve or think about a problem. Cognitive outcomes can be measured with various types of items, such as short essay, sentence-completion, and multiple choice. These items are often categorized into open-ended and closed-ended response formats. Closed-ended response formats, such as multiple choice, true-false, and matching items, are those in which the item choices or responses are provided for the respondent. Open-ended formats, such as short answer and essay, require the respondent to construct their own answer. These items are also called constructed-response items.

Closed-ended Items

Multiple choice items are typically considered the most versatile type of item. Although these types of items are often criticized for only measuring rote memorization, multiple choice items can be constructed in such a way as to tap into higher level cognitive processes, such as synthesis and analysis of information. Items that ask respondents to interpret tables or graphs, make comparisons, or identify similarities or differences, focus on higher level cognitive processes. Another advantage of multiple choice items is that they can provide the test administrator with helpful diagnostic information about respondents' misunderstandings of the subject matter. To do this, incorrect answer choices (known as *distractors*) must be written to include common errors or misconceptions.

True-false items are also common cognitive because they are considered easier to write than multiple choice items. This is because true-false items do not contain distractors. However, true-false items are more susceptible to being influenced by guessing than multiple choice items. Matching items, which typically involve a list or set of related responses, can be used to measure a wide range of skills (from simple to complex knowledge) and are often used to assess knowledge of specific facts.

Open-ended Items

Open-ended or constructed response items require the respondent to formulate and write an answer to a prompt. Open-ended items include short answer/essay and sentence completion items. An advantage of these items is that they are not susceptible to guessing. However, since these items allow more freedom for respondents to answer the question, they are not as easy to score as closed-ended items. These items are typically scored using a rubric or checklist. It is also important to note that one should examine inter-rater agreement or reliability when using a rubric. Inter-rater reliability is not addressed within this handbook. However, a reference to this issue is provided at the end of the guidebook.

Non-Cognitive or Attitudinal Items

Non-cognitive or attitudinal items measure the preferences and/or interests of respondents. An attitude typically refers to how one feels toward an object, idea, or person. Other non-cognitive outcomes may include beliefs and values. There are no right or wrong answers to these type of items.

While item writers do not need to be concerned about respondents guessing the correct answers (as in cognitive items), they should be concerned about other issues, such as social desirability. Social desirability refers to a response pattern in which respondents select answers that they believe are socially desirable options. For example, if an assessment practitioner aimed to measure "risky" or "illegal" behavior, students may be hesitant to provide truthful responses. Social desirability can, at least in part, be addressed by ensuring confidentiality or anonymity to participants.

Similar to cognitive items, attitudes can be assessed in several ways. These items can be open-ended (e.g., short answer) or closed-ended (e.g., ranking scale or multiple response). Closed-ended item types are described in greater detail below.

Closed-ended Items

Type of Item	**Description**
Checklists	Can be used to obtain a great deal of information at one time. A checklist can be used to indicate the presence or absence of something of interest.
Multiple-response Items	These items present a problem and offer the examinee response options. Typically these items allow respondents to select more than one response. (e.g., "select all that apply").
Ranking Scales	Used to rank-order things as they relate to one another. For example, a respondent orders a list of beverages according to the preferences. Should be limited to no more than five items to avoid confusion and misnumbering by respondents.
Likert-type Scales	Composed of items that ask respondents to rate a statement using a scale, such as 1 = strongly agree to 5 = strongly agree. The items may yield a total score or several subscale scores.

General Guidelines for Item Writing

The following guidelines apply to the writing of cognitive or attitudinal items. The item format is specified where appropriate.

Cognitive Items

- Offer 3-4 well-developed answer choices for multiple choice items. The aim is to provide variability in responses and to include plausible distractors.
- Distractors should **not** be easily identified as wrong choices.
- Use **boldface** to emphasize negative wording.
- Make answer choices brief, not repetitive.
- Avoid the use of *all of the above, none of the above,* or a combinations such as *A and B* options. It is tempting to use them because they are easy to write, but there are several reasons for avoiding their use:
    - Students with partial knowledge of the question may be able to answer the question correctly by process of elimination. For example, a student may know that two out of the three response options are correct, causing them to correctly select *all of the above*.
    - These items may make it more difficult to discriminate between those students who fully know the subject matter and those who do not, which may lower reliability.
- For matching items, provide more response options than items in the list (e.g., offer 10 response options for a list of seven items). This decreases the likelihood of participants getting an answer correct through the process of elimination.
- List response options in a set order. Words can be listed in alphabetical order; dates and numbers can be arranged in either ascending or descending order. This makes it easier for a respondent to search for the correct answer.
- Make sure do not give hints about the correct answer to other items on the instrument.

Attitudinal Items

- Avoid "loading" the questions by inadvertently incorporating your own opinions into the items.
- Avoid statements that are factual or capable of being interpreted as such.
- Avoid statements that are likely to be endorsed by almost everyone or almost no one.
- Statements should be clearly written -- avoid jargon, colloquialisms, etc.
- Each item should focus on one idea.
- Items should be concise. Try to avoid the words "if" or "because", which complicates the sentence.

Common Mistakes when Writing Items

Double-barreled items

These are items that express two or more ideas.

- Example: “I like to exercise at the gym at least 3-5 times per week.” (i.e., Is the item addressing whether respondents like to exercise at the gym or how often they like to exercise?
- Reworded: I like to exercise at the gym.
  I like to exercise at least 3-5 times per week.

Items that give clues to the correct answer in the wording

- Example: When the two main characters went outside they tossed around **an** ___________________.
  a. baseball
  b. tomato
  c. apple
  d. football
- Reworded: When the two main characters went outside they tossed around **a(n)** ___________________.
  OR
- When the two main characters went outside they tossed around **the** ___________________**.**

Loaded questions

These questions inadvertently incorporate your own opinion into the item.

- Example: Is there any reason to keep this program?
- Reworded: Does this program offer material not obtained in other courses?

RELIABILITY

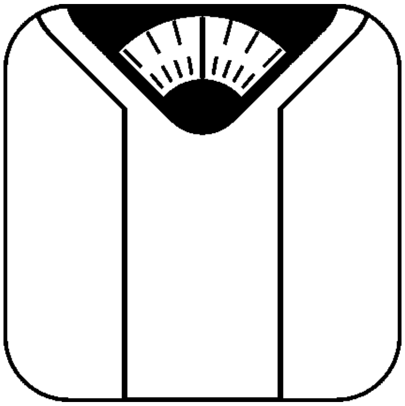
There are various approaches for examining “reliability-like” coefficients within measurement literature. Our discussion focuses on Classical Test Theory (CTT) since this is perhaps the most widely used approach among assessment practitioners. We do not address measurement error from the perspective of generalizability theory or item response theory within this guide.

We begin by providing a conceptual overview of reliability which is followed by a summary of how it is assessed. This section concludes by discussing some practical considerations in reliability estimation.

Conceptual Overview

Reliability is concerned about the consistency of scores. Imagine an individual who repeatedly stands on a bathroom scale within a few minutes and records each number. Further imagine that each time this person stood on the scale a random number appeared.

We may obtain something like the following numbers: 10, 40, 95, 120, 123, 140, 150, and 205.



A similar problem exists when we think about assessment. Imagine that Student 1 took a test and received a score of 85. Ideally, if Student 1 were to take the same test under the same conditions their new score should be close to 85. If these scores were widely inconsistent across independent replications of testing then we would should be concerned about using these scores to make judgments about Student

Most measurement theorists believe that reliability (i.e. consistency) is a necessary but insufficient condition for measurement. *If scores were entirely inconsistent then they are random. If scores are random, then nothing is being measured.*

Conceptual Overview of Classical Test Theory

Classical Test Theory (CTT) is an axiomatic system that allows us to estimate reliability in a population if in fact particular assumptions are true. According to CTT each person's score can be described by the following equation:

$$X = T + E \tag{1}$$

In this equation X refers to a person's observed score. For Student 1 the observed score refers to the 85 on the test described above. Their score of 85 reflects the composite of a true score (i.e., T in the equation) and error (i.e., E in the equation).

The term "true" in CTT can be misleading. *A true score does not reflect the true or actual level of an attribute*. In other words, if this were a math test Student 1's true score does not indicate their actual math ability. *Instead the true score is something like an average, or more precisely an expected value, across repeated replications of administering the same test to Student 1.* Error is by definition random.

By conducting simple manipulations of equation 1 we can see that a true score can also be conceived as the difference between an observed score and error (i.e., T = X -E) and that error can also be conceived as the difference between a true score and observed score (i.e., E = T - X).

Since we are interested in populations, as opposed to a particular student, we apply the following equation:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{2}$$

Equation 2 indicates that observed score variance (σ_X^2) is composed of true score variance(σ_T^2) plus error variance (σ_E^2).

Reliability Defined

From equation 2 we can derive different, though mathematically equivalent, ways to conceptualize reliability according to CTT. Two equivalent ways to conceptualize reliability are provided in the Table below.

Reliability Equation	Interpretation
$\text{Reliability} = \dfrac{\sigma_T^2}{\sigma_X^2}$	Ratio of true score variance to observed score variance. In other words, it is <u>the amount of observed score variance that we can attribute to true score variance</u>. A value of .80 indicates that we can attribute 80% of the observed score variance to true score variance.
$\text{Reliability} = 1 - \dfrac{\sigma_E^2}{\sigma_X^2}$	This is the <u>absence of error variance</u>. In other words, it is the amount of observed score variance that is systematic as opposed to random. A reliability estimate of .80 indicates that 20% of the observed score variance may be attributed to error.

Reliability estimates range from 0 to 1. A value of 0 indicates that all of the variance may be attributed to error. A value of 1 indicates that all of the variance can be attributed to true score variation.

Important Assumptions

As previously mentioned, CTT consists of a set of axioms, along with assumptions, that allow us to estimate reliability. It is important to know when such assumptions may be violated to be able to evaluate the appropriateness of reliability estimates in different situations.

The table below provides an overview of important axioms and assumptions used in CTT. This table also describes a scenario in which each assumption/axiom may be violated.

Equation	Description	Example of Violation
X = T + E	Each score is composed of a true score and random error.	This cannot be tested.
ε(X) = T	A true score is an expected value across repeated replications of a measurement procedure. True scores are assumed to be constant between repeated replications.	This is the definition of a true score. True scores should not change across replications – this could be violated if students developed between testing occasion.
$\rho_{ET} = 0$	The correlation between error and true scores is zero.	Children have assigned seating according to ability. A disturbance in the back of the room interferes with testing.
$\rho_{E1E2} = 0$	Error between replications should not be correlated.	Two tests are administered at the end of a long battery of exams on separate occasions. Error on both may be correlated due to fatigue.
$\rho_{E1\mathrm{T}2} = 0$	Error on test 1 should not correlate with true scores on test 2.	Similar situations as assumption 3.

Note. X = observed score; T = true score; E = error; ε = expected value; ρ = population correlation.

Sources of Error

Error is basically an indication of score inconsistency. Error is assumed to be random. When thinking about a set of scores, there are various ways in which these scores could be inconsistent. For example, scores may be inconsistent over time, across different forms of a test, or across a set of items written to measure the same attribute. Test-retest reliability is concerned about consistency over time, equivalent forms is concerned about consistency across two versions of a test, and internal consistency is concerned about consistency across a set of items. Each of these are reviewed in turn.

Test-Retest

As previously mentioned, test-retest reliability is concerned with the consistency of scores across two points in time. This reliability estimate is obtained by examining the correlation between both sets of scores across each point in time (i.e. scores at Time 1 with scores at Time 2). This is often referred to as a stability coefficient since it indicates the extent to which the relative position of students tends to be stable across measurement occasions. The following is a list of issues to consider when using this method:

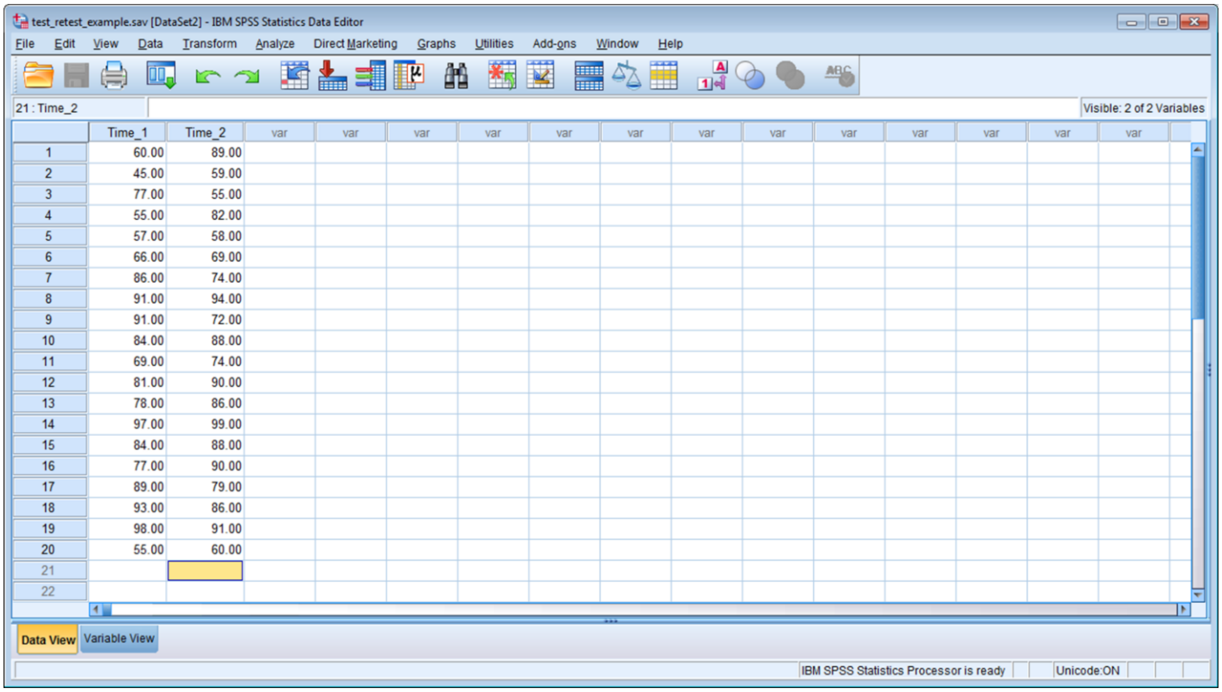
- Test-retest reliability is particularly important in situations in which we want to take scores at Time 1 and use them at Time 2. For example, if students complete a math placement test in the early summer and these scores are used to place students into particular courses in August then the scores should be consistent across this interval of time.
- Test-retest reliability should not be estimated before and after an experimental intervention. This includes any situation in which students have received instruction. We do not expect scores to be consistent in this situation because we would like for the intervention to be effective.
- Test-retest reliability is inappropriate in situations where we expect an attribute to change over time such as moods or other psychological states. Participant sensitization (e.g. they become aware that you are examining test anxiety) and practice effects (e.g. students score increase due to familiarity with the instrument) may also be a concern when administering the same instrument twice.
- Selecting an ideal interval of time between testing is perhaps the most challenging issue facing individuals using this technique. Unfortunately there are no simple answers to this question. In general, one should consider whether it is reasonable for scores to be stable over a given time interval and how you plan to use the scores to make decisions.
- There are no clear guidelines for acceptable test-retest correlations since these values depend on a number of contextual factors (e.g. type of attribute, length of interval, etc.).

SPSS Test-Retest

In this example, we have 20 students who were administered a math placement test in May. These students were administered the same test in August.

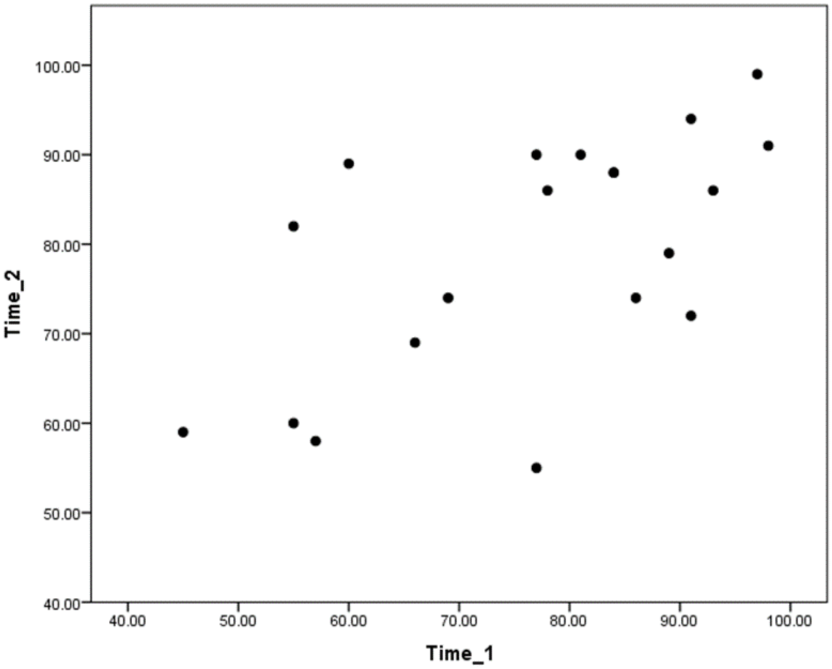
This data would be displayed as follows in SPSS. In this file, each row is a person. Variables, which in this case represent scores at Time 1 and Time 2, are columns.

Prior to estimating our reliability coefficient, we may wish to examine our data using a scatterplot. This is provided below.

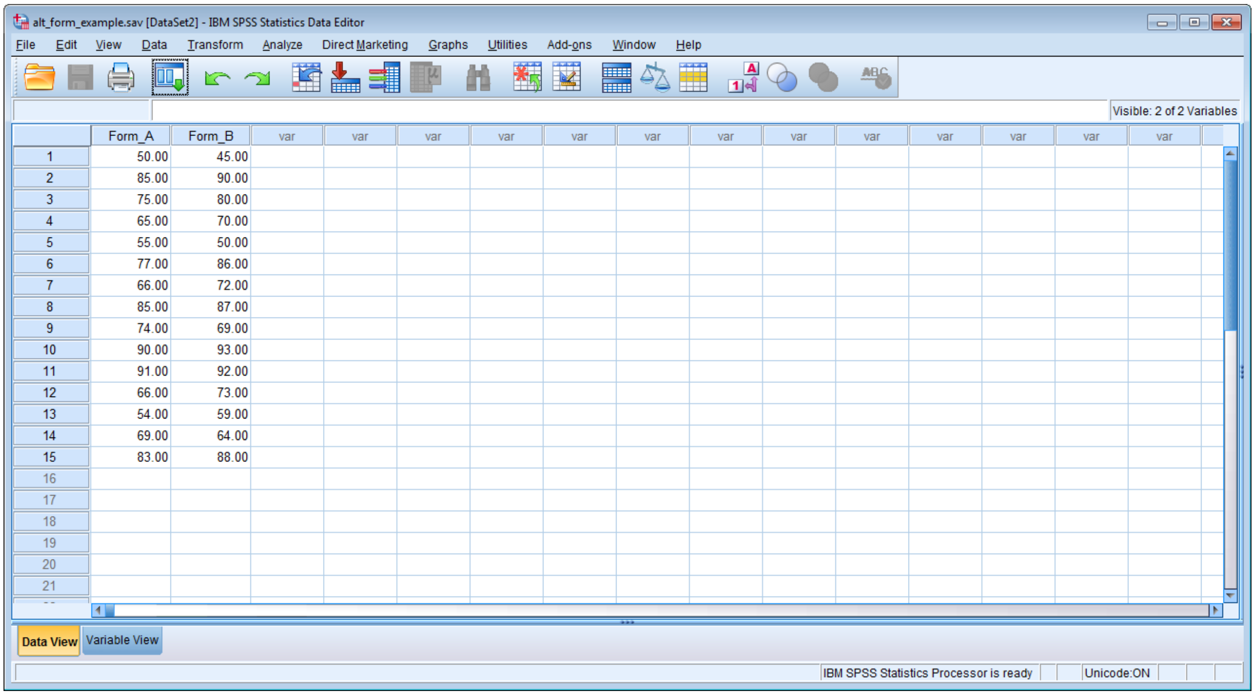

test_retest_example.sav [DataSet2] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
21 : Time_2
Visible: 2 of 2 Variables
var
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

	Time_1	Time_2
1	60.00	89.00
2	45.00	59.00
3	77.00	55.00
4	55.00	82.00
5	57.00	58.00
6	66.00	69.00
7	86.00	74.00
8	91.00	94.00
9	91.00	72.00
10	84.00	88.00
11	69.00	74.00
12	81.00	90.00
13	78.00	86.00
14	97.00	99.00
15	84.00	88.00
16	77.00	90.00
17	89.00	79.00
18	93.00	86.00
19	98.00	91.00
20	55.00	60.00
21		
22		

Scatterplot of Student Scores at Time 1 and Time 2



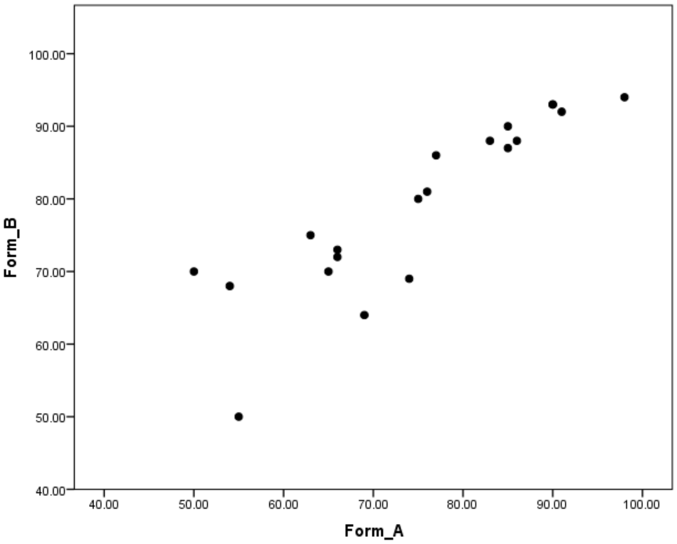
The scatterplot displays each score across both points in time. This plot suggests that there is a tendency for students who score high at Time 1 to also score high at Time 2. Similarly, students who score low at Time 1 tend to score low at Time 2.


alt_form_example.sav [DataSet2] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
Visible: 2 of 2 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

	Form_A	Form_B	var	var	var	var	var	var	var	var	var	var	var	var	var
1	50.00	45.00													
2	85.00	90.00													
3	75.00	80.00													
4	65.00	70.00													
5	55.00	50.00													
6	77.00	86.00													
7	66.00	72.00													
8	85.00	87.00													
9	74.00	69.00													
10	90.00	93.00													
11	91.00	92.00													
12	66.00	73.00													
13	54.00	59.00													
14	69.00	64.00													
15	83.00	88.00													
16															
17															
18															
19															
20															
21															

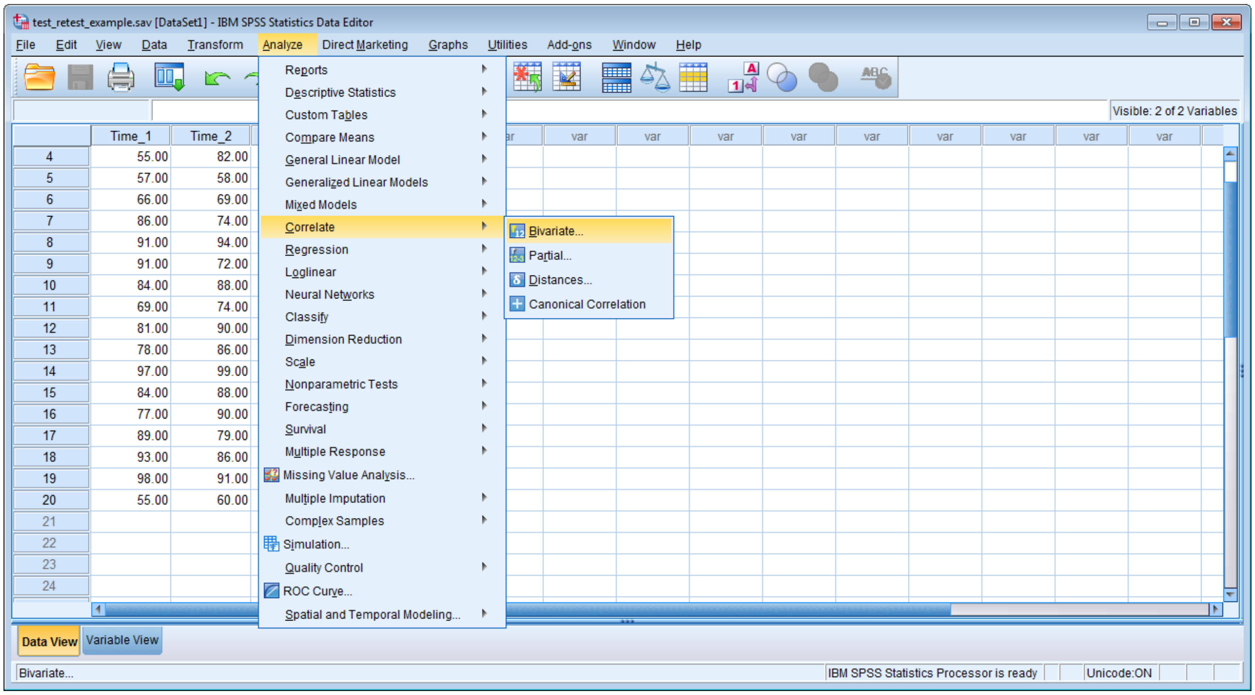
Before examining the correlation between each form we may wish to examine the scatterplot. The scatterplot indicates that there is a tendency for students who score high on Form A to also score high on Form B. Similarly, students who have low scores on Form A also tend to have low scores on Form B.

Scatterplot of Student Scores for Form A and Form B

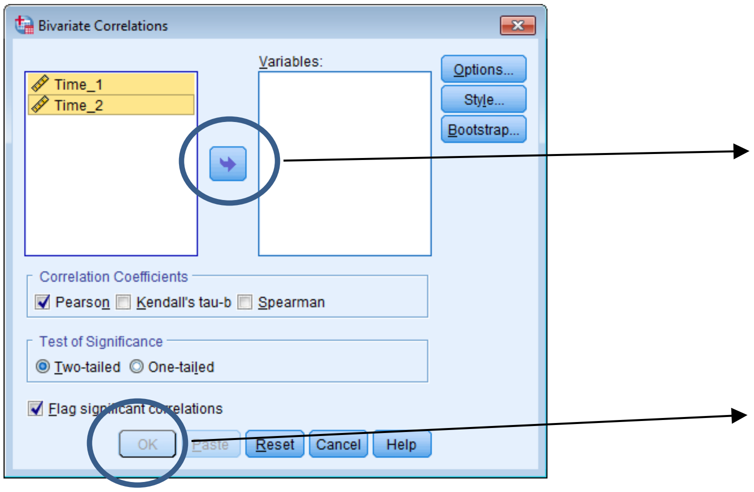


To obtain the test-retest reliability estimate we need to obtain the correlation between scores at Time 1 and scores at Time 2. The steps in SPSS are described as follows:

First go to “Analyze” -→ “Correlate” → Bivariate



	Time_1	Time_2
4	55.00	82.00
5	57.00	58.00
6	66.00	69.00
7	86.00	74.00
8	91.00	94.00
9	91.00	72.00
10	84.00	88.00
11	69.00	74.00
12	81.00	90.00
13	78.00	86.00
14	97.00	99.00
15	84.00	88.00
16	77.00	90.00
17	89.00	79.00
18	93.00	86.00
19	98.00	91.00
20	55.00	60.00
21		
22		
23		
24		



Correlation Coefficients

☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance

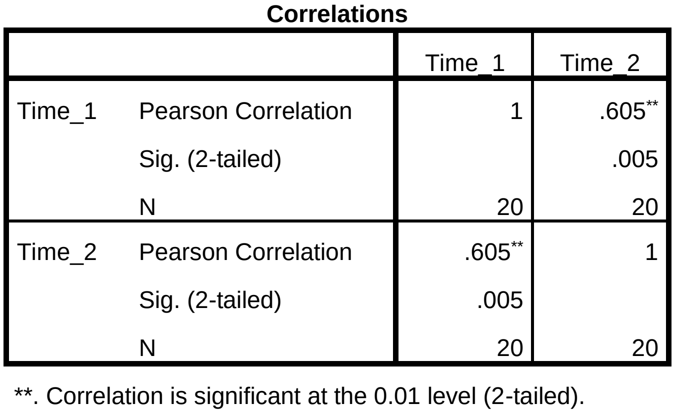
● Two-tailed ○ One-tailed

☑ Flag significant correlations

Highlight Time 1 and Time 2 and then move them to the “Variables” box.

After completing the step above the “OK” button will change colors. Click “OK.”

After clicking “OK” you will obtain the following output.



Correlations

		Time_1	Time_2
Time_1	Pearson Correlation	1	.605**
	Sig. (2-tailed)		.005
	N	20	20
Time_2	Pearson Correlation	.605**	1
	Sig. (2-tailed)	.005	
	N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

In this example, our test-retest correlation was .605. Though this value was statistically significant, since the *p*-value (i.e., sig. 2-tailed) is .005, the magnitude of the reliability estimate is relatively small for making placement decisions. There are no strict rules-of-thumb for such decisions. However if our estimate was closer to .80 or .90 we would have more confidence that the scores in May could be used to make placement decisions in August.

Such inconsistencies may occur for various reasons. For example, it is possible that students simply fail to remember important math skills if they take a break during the summer. Irrespective of the reasons for such inconsistencies, we would recommend for the students to take the placement test in August since these scores may be a better indication of skills before the fall semester.

Alternate Forms

There are situations in which a practitioner may wish to create two versions of a test or they may be interested in the consistency of scores between two versions of a test that have already been created. In such instances, scores should not depend upon which test an examinee happens to receive. For example, the score of Student 1 should not depend on which form of the test they happened to receive.

Similar to the test-retest estimate, this source of error is examined by correlating scores between Form A and Form B among the same sample of students. This reliability estimate is often referred to as a “coefficient of equivalence” since it examines the extent to which the relative standing, or position, of students is consistent between two versions of a test. The following is a list of issues to consider when using this method:

- This reliability estimate is important when we would like to be able to substitute one test for another test.
- Controlling for potential fatigue is an issue with this technique. For example, it is unlikely that students will be able to take both tests on the same day due to fatigue.

It is generally recommended to administer both forms as closely as possible without evoking fatigue.

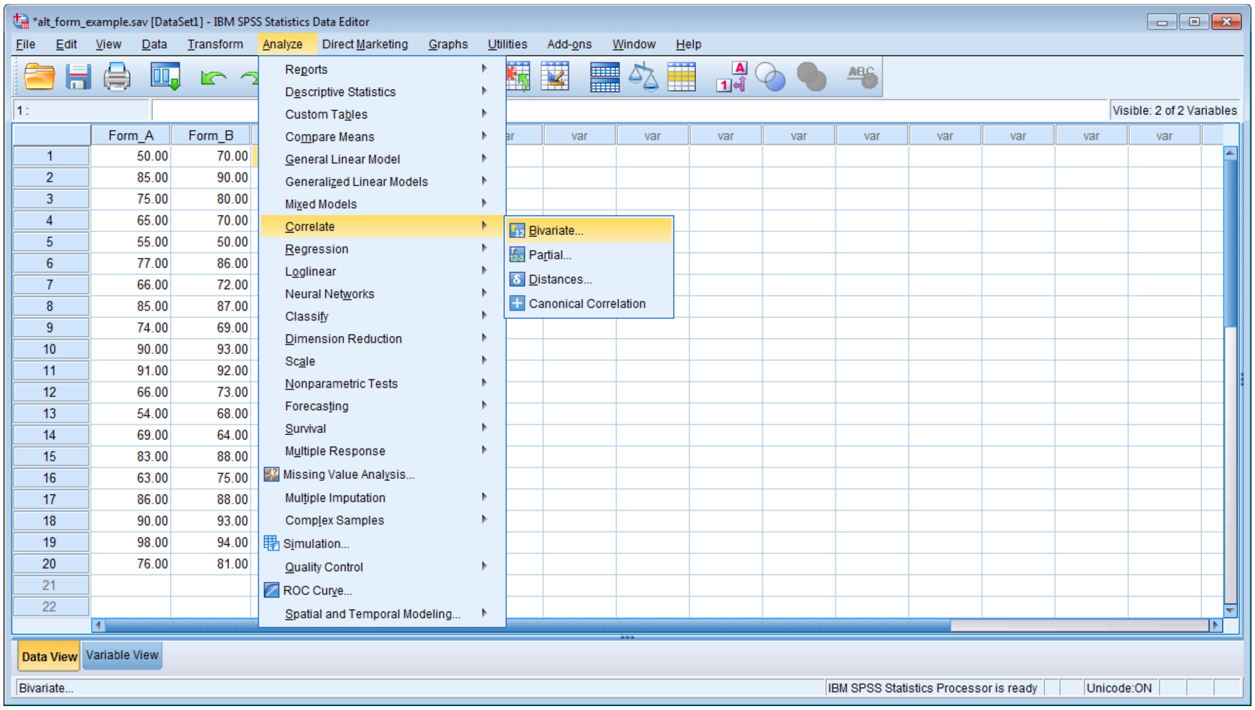
- In some cases, simply being exposed to the content of test can increase scores on subsequent testing (i.e., practice effect). It is therefore recommended to counterbalance the administration of each form. In other words, half of the participants are assigned to complete Form A followed by Form B and the other half are assigned to complete Form B followed by Form A.
- A lack of consistency between forms may be due to differences in: (a) difficulty, (b) content, (c) cognitive complexity, and/or (d) issues with fatigue.
- We would like to see correlations around .80 to .90.

SPSS Alternate Forms

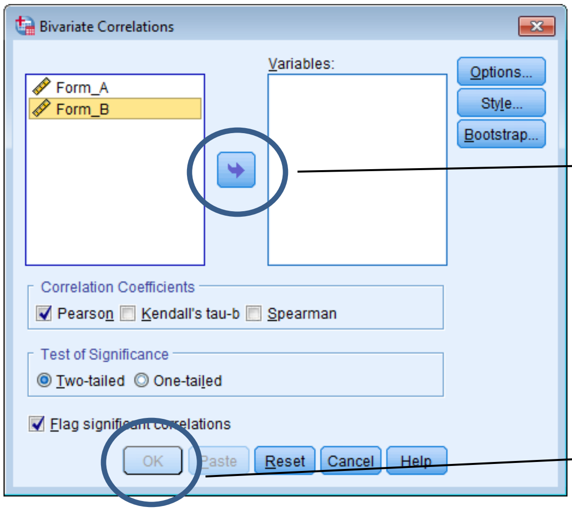
An assessment practitioner is interested in a math placement test. Students are usually assigned Form A; however, recently a testing company released Form B. The practitioner is interested in whether students would obtain similar scores across each form. The practitioner conducts a small study wherein 20 students were administered Form A and Form B.

The data setup, and procedures, for this scenario are the same way as what was done for test-retest reliability. Students constitute rows and the variables (i.e., form) are depicted as columns.

Go to “Analyze” → “Correlate” → “Bivariate”



	Form_A	Form_B
1	50.00	70.00
2	85.00	90.00
3	75.00	80.00
4	65.00	70.00
5	55.00	50.00
6	77.00	86.00
7	66.00	72.00
8	85.00	87.00
9	74.00	69.00
10	90.00	93.00
11	91.00	92.00
12	66.00	73.00
13	54.00	68.00
14	69.00	64.00
15	83.00	88.00
16	63.00	75.00
17	86.00	88.00
18	90.00	93.00
19	98.00	94.00
20	76.00	81.00
21		
22		



Correlation Coefficients

☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance

● Two-tailed ○ One-tailed

☑ Flag significant correlations

Highlight Time 1 and Time 2 and then move them to the "Variables" box.

After completing the step above the "OK" button will change colors. Click "OK."

After clicking "OK" you will obtain the following output.

Correlations

		Form_A	Form_B
Form_A	Pearson Correlation	1	.953**
	Sig. (2-tailed)		.000
	N	15	15
Form_B	Pearson Correlation	.953**	1
	Sig. (2-tailed)	.000	
	N	15	15

**. Correlation is significant at the 0.01 level (2-tailed).

In this case, the correlation between each form was .953 (*p* < .001). The magnitude of this correlation is large and nearly close to 1 which would indicate a perfect correlation. This study supports the position that the relative standing of students on math placement test does not depend on which form they happen to receive.

Internal Consistency

Items written to assess the same attribute should be correlated. For example, students who tend to get Item 1 correct should also have a tendency to get Item 2 correct. With respect to attitudinal items, students who endorse (e.g., "agree") with a statement about their satisfaction with life should agree to similar statements about their life satisfaction. Internal consistency is basically an indication of "interrelatedness" among a set of items.

Coefficient alpha is perhaps the most frequently reported estimate of internal consistency for attitudinal items. KR-20 is a similar estimate reported for dichotomous items (e.g. right/wrong, yes/no, etc.). SPSS uses the label "coefficient alpha" for both estimates.

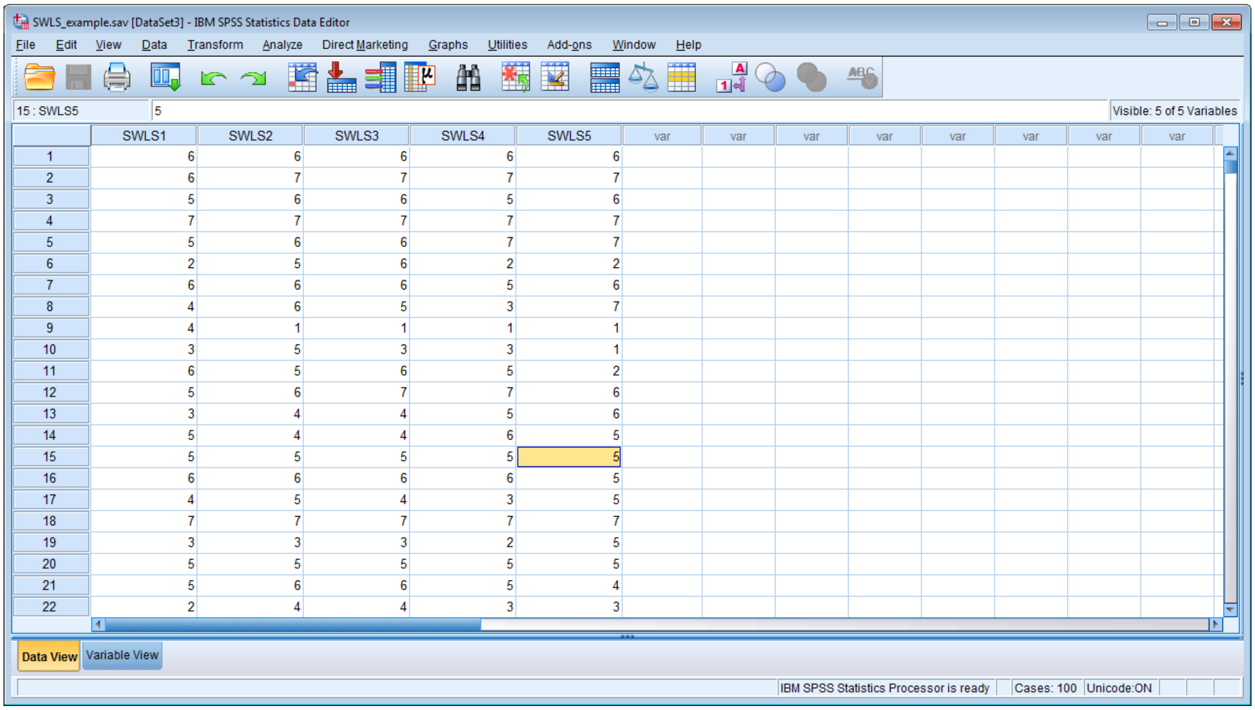
- Alpha is a function of "interrelatedness" and the number of items. Consequently, alpha can be high simply because there are a lot items on an instrument.
- Alpha is often interpreted as a lower-bound estimate of reliability. Alpha can actually be an overestimate of reliability in some situations (e.g. mistakes on an item are correlated). Additionally, alpha fails to consider other sources of error, such as time.
- Alpha does *not* indicate that the items are measuring the same thing. Instead, alpha *assumes* that items are measuring the same thing. Items can be correlated for multiple reasons, only one of which is due to measuring the same attribute.
- Alpha should not be reported when a test is speeded. This will result in an overestimate of reliability. With speeded tests, it is more appropriate to split the test in two and correlate each half. This correlation will need to be corrected using the Spearman-Brown prophecy formula to estimate reliability across the entire test (located in more technical books).
- Values below .70 are generally acceptable for research purposes. In high-stakes testing contexts, these values may need to much higher such as around .80 or .90.

SPSS Alpha Coefficient and Item Analysis

Here we provide an overview of how to obtain an alpha coefficient in SPSS. In situations in which you have dichotomous data, you indicate in SPSS that the variable is categorical as opposed to being on a scale. SPSS reports KR-20 as coefficient alpha. This section will also review some basic item analyses that one may choose to conduct when examining alpha.

In this example, we are using real data obtained from a sample of 652 undergraduate students at a large university in the Midwest. These students were administered 5 items from a satisfaction with life scale as part of a larger study. Each item is scored on a Likert-type scale ranging from 1 = strongly disagree to 7 = strongly agree.

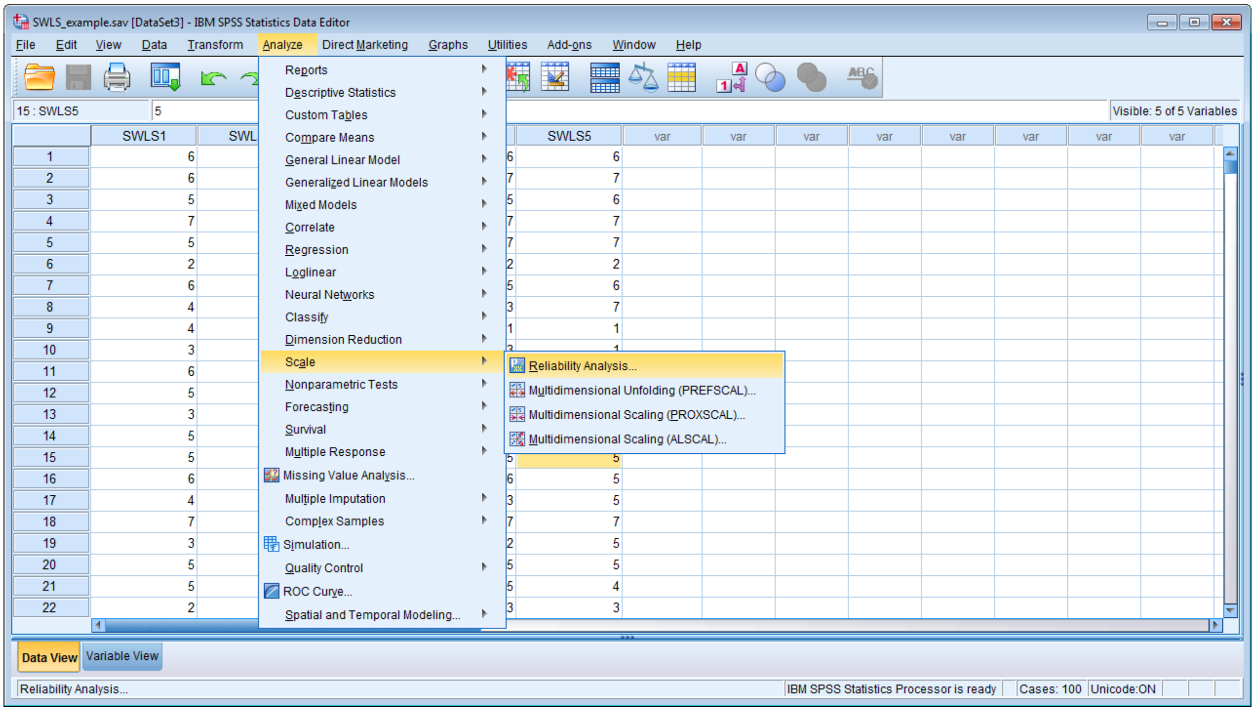
The SPSS data file is set up in the same way as previous examples. Students constitute rows and their responses to the five items are depicted in the columns. SWLS1 is response to item 1, SWLS2, is response to item 2, etc.

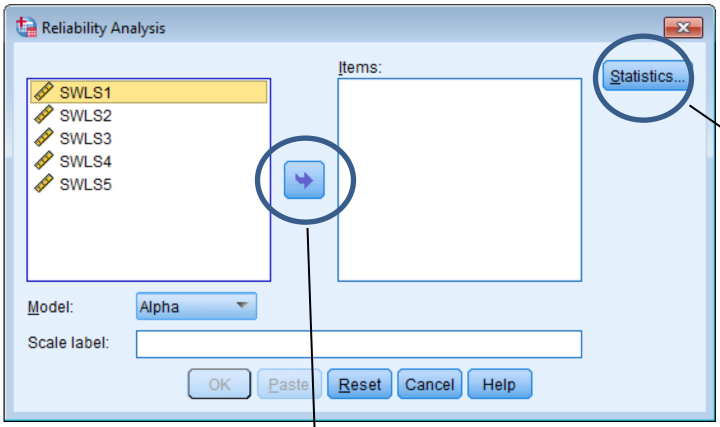

SWLS_example.sav [DataSet3] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
15 : SWLS5
5
Visible: 5 of 5 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Cases: 100
Unicode:ON

	SWLS1	SWLS2	SWLS3	SWLS4	SWLS5	var	var	var	var	var	var	var	var
1	6	6	6	6	6								
2	6	7	7	7	7								
3	5	6	6	5	6								
4	7	7	7	7	7								
5	5	6	6	7	7								
6	2	5	6	2	2								
7	6	6	6	5	6								
8	4	6	5	3	7								
9	4	1	1	1	1								
10	3	5	3	3	1								
11	6	5	6	5	2								
12	5	6	7	7	6								
13	3	4	4	5	6								
14	5	4	4	6	5								
15	5	5	5	5	5								
16	6	6	6	6	5								
17	4	5	4	3	5								
18	7	7	7	7	7								
19	3	3	3	2	5								
20	5	5	5	5	5								
21	5	6	6	5	4								
22	2	4	4	3	3								

To obtain alpha:

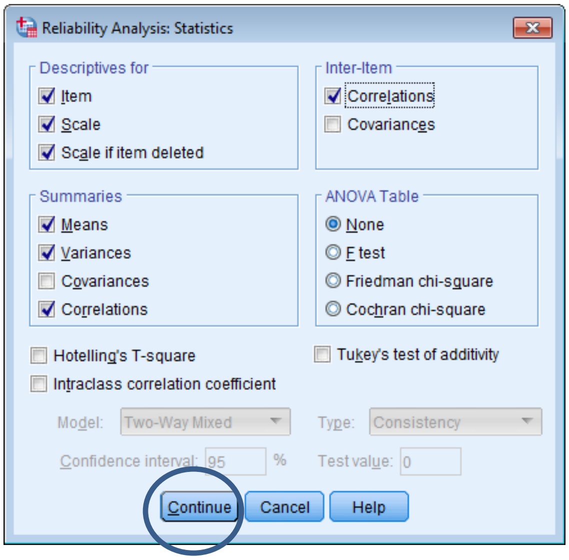
Click “Analyze” → “Scale” → “Reliability Analysis”


SWLS_example.sav [DataSet3] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Custom Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
ROC Curve...
Spatial and Temporal Modeling...
Reliability Analysis...
Multidimensional Unfolding (PREFSCAL)...
Multidimensional Scaling (PROXSCAL)...
Multidimensional Scaling (ALSCAL)...
15 : SWLS5
5
Visible: 5 of 5 Variables
Data View
Variable View
Reliability Analysis...
IBM SPSS Statistics Processor is ready
Cases: 100
Unicode:ON



Click “Statistics” to open up a new box.

Move each variable into the “Items” box.



Descriptives for
☑ Item
☑ Scale
☑ Scale if item deleted

Inter-Item
☑ Correlations
☐ Covariances

Summaries
☑ Means
☑ Variances
☐ Covariances
☑ Correlations

ANOVA Table
● None
○ F test
○ Friedman chi-square
○ Cochran chi-square

☐ Hotelling's T-square
☐ Tukey's test of additivity
☐ Intraclass correlation coefficient

Model: Two-Way Mixed
Type: Consistency
Confidence interval: 95 %
Test value: 0

Before clicking “Continue” put a check in any box for which you like to see output. In this case, we checked descriptives for item, scale, and scale if item deleted. We also checked the box to obtain inter-item correlations and the overall mean, variance, and average inter-item correlation. After clicking “continue” a new box will appear. Click “OK” to obtain the output.

A brief review of inter-item correlations, alpha, and some of the item-total statistics will be provided.

Inter-Item Correlation Matrix

	SWLS1	SWLS2	SWLS3	SWLS4	SWLS5
SWLS1	1.000	.676	.611	.549	.358
SWLS2	.676	1.000	.674	.579	.380
SWLS3	.611	.674	1.000	.661	.446
SWLS4	.549	.579	.661	1.000	.424
SWLS5	.358	.380	.446	.424	1.000

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	5.291	4.834	5.685	.851	1.176	.107	5
Item Variances	2.147	1.711	3.440	1.729	2.010	.542	5
Inter-Item Correlations	.536	.358	.676	.318	1.887	.015	5

Here we have the inter-item correlation matrix as well as some summary information. Correlations range from .358 to .676 with an average inter-item correlation of .536. SWLS5 has slightly smaller correlations than other items. For example, the correlation between SWLS5 is .358 and .380 with items SWLS1 and SWLS2 respectively.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.836	.852	5

The circled value is our alpha coefficient, which is .836. We will now review some item-total statistics to further evaluate the quality of each item (see next table).

The corrected item-total correlations provides the correlation between each item and the total score after removing the item from the total. In CTT, this correlation is referred to as item discrimination. Conceptually, these correlations reflect a tendency for people who endorse an item to obtain higher scores on the overall test. There are no standard guidelines for interpreting these values; however, values below .20 may suggest that an item is problematic. In our case, the values range from .475 to .667.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
SWLS1	21.33	22.165	.667	.515	.796
SWLS2	20.98	21.952	.710	.577	.786
SWLS3	20.77	21.591	.749	.593	.776
SWLS4	21.11	21.329	.682	.496	.791
SWLS5	21.62	20.913	.475	.235	.868

The “Cronbach’s Alpha if Item Deleted” provides the value of our new alpha coefficient if we deleted a particular item from the scale. For example, our original alpha was .836 and if we removed SWLS1 our new alpha would be .796. This is expected since removing items from a scale should decrease alpha.

SWLS5 appears to be more problematic than other items. If we removed SWLS5 we see that alpha slightly increases from .836 to .868. Though this this may not be viewed as a huge increase, the item appears to be contributing to unreliability. The content of this item should be examined to investigate why SWLS5 appears to be functioning differently from other items.

Selecting a Reliability Estimate

Selecting an appropriate reliability estimate is primarily determined by what you are attempting to do with the scores. As a practitioner, you must think through what source of inconsistency is likely to be an issue with respect to how you propose to use an instrument. For example, if no alternate forms exist then this source is simply a non-issue. Internal consistency should probably always be reported since this is relatively easy to accomplish and is arguably a concern in most situations. Test-retest may also be unreasonable in some situations, such as when an attribute is expected to rapidly change between testing.

Additional Issues for Consideration

- *Reliability is a property of scores in CTT. It is not a property of a test. Consequently, it is inappropriate to indicate that a test is reliable or unreliable. You should always collect this information on your sample.*
- Reliability estimates are population dependent in CTT. Reliability estimates can change across different populations; thus when speaking about reliability one should clarify the population of interest.

- The same issues which impact correlations can also influence reliability. For example, correlations will be lower when the range is restricted. Reliability will tend to be lower in homogenous samples or samples with highly similar scores.

VALIDITY

Validity is arguably one of the most misunderstood concepts in educational and psychological measurement. Part of this confusion may result from the fact that the term is used differently both between and within disciplines. For example, in logic the term validity is used to describe arguments of a particular form (i.e., one in which the conclusion follows from each premise). In experimental design internal and external validity are used to indicate the extent to which one may reasonably infer the presence of an effect (i.e., internal) and the extent to which such an effect fails to change across people, settings, and so forth (i.e., external).

Our interest in the concept of validity is delimited to educational and psychological measurement. It should be noted however, that the concept of validity has changed over time even within this field and that measurement specialists disagree about several nuances regarding this concept. Our discussion of validity draws from the most recent Standards for Educational and Psychological Testing (2014) since this book is perhaps the best representation of a "consensus" view within the field of educational and psychological measurement.

Definition and Displacing Common Misconceptions

According to the Standards for Educational and Psychological Testing (2014), there is a distinction between "validity" and "validation." *Validity is "the degree to which evidence and theory support interpretations of test scores for proposed uses of tests"* (p. 11). *Validation on the other hand, "can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use"* (p. 11).

Five basic points derive from these definitions.

1. Tests are not valid or invalid

Strictly speaking, it is inappropriate to say that a test is valid. Scores may be interpreted in multiple ways. For example, one practitioner may view a set of scores as an indication of differences in personality whereas another practitioner could view these scores as an indication of behavioral preferences.

Tests are also used to make various decisions about students. For example, the personality test could be used to make decisions about student interest in specific disciplines, facilitate theoretical research, or to advise students about the adequacy of potential marital partners.

Broadly claiming that "a test is valid" fails to consider the fact that scores may be interpreted and used to achieve multiple aims; evidence for one aim does not necessarily justify using the test to achieve a different aim.

2. Validity is a matter of degree

Evidence and theoretical support for a particular interpretation or use of a test may change over time. For example, in 1995 existing evidence may suggest it is appropriate to use a test place students into a particular program. Subsequent research however, may suggest that these initial studies were misleading.

Since evidence and theoretical support change over time, it is technically inappropriate to say that a test is valid. Instead, *interpretations for proposed uses of tests are more or less valid*, which is a function of evidential support. Unfortunately, evidence is at times deceptive in that it leads to inappropriate conclusions. If our judgments about validity are based upon fallible evidence, then is inappropriate to conclude that a test is either valid or invalid.

3. Validity is a unitary concept

Various research methods textbooks, particularly those discussing validity in the context of measurement, often promote the idea that there are different types of validity. These types generally include: (a) content, (b) criterion, and (c) construct.

The idea that there are different types of validity has been rejected by an overwhelming majority of measurement theorists since the 1970's. Instead of "types" of validity, the Standards of Educational and Psychological Testing (2014) describe different sources of evidence. Validity is a unitary concept. In other words, it is inappropriate to describe different types of validity.

4. Each interpretation must be supported

As previously mentioned, scores on a test may be interpreted in various ways. Researchers/ practitioners often disagree about the best way to interpret a set of scores. Each interpretation must be logically and empirically examined.

5. Each use of a test must be supported

Tests may be used to achieve multiple purposes (e.g. theory testing, selection, placement, prediction, etc.). Thus, evidence that a test may be used to achieve one purpose does not imply it may be used to achieve a different purpose. Each proposed use of a test should be evaluated.

Threats to Validity

There are two fundamental threats to validity, which include construct-underrepresentation and construct-irrelevant variance. Validation may be viewed, at least in part, as the process of examining the extent to which each of these threats limit specific interpretations/uses of a test.

Construct-underrepresentation indicates that a test is "too narrow" since it is missing something important. To provide a simple example, a teacher may interpret a math test as indicating differences in student's knowledge of addition and subtraction. This would be a threat, for example, if the teacher included addition problems but no subtraction problems. It may also be a problem if the teacher failed to represent the "breadth" of subtraction problems actually covered in class (e.g. all the problems on the test required students to subtract single digits).

Construct-irrelevant variance indicates that a test is "too broad" since the differences in the scores are inadvertently influenced by something we do not want. This can manifest in several ways. For example, it is possible that the correlation between leadership and a personality trait is in part influenced by each variable being collected in a similar manner (e.g. a method effect due to relying on self-report). In such a situation, one may choose to also measure each variable using a different technique, such as observations or friend reports, to examine whether variables that presumably measure the same trait are more correlated than variables measuring different traits (i.e. multi-trait multi-method matrix).

With respect to the math example, construct-irrelevant variance could be an issue if a teacher primarily included "word problems" requiring a high reading level. In this case, it is possible that the differences in math scores are unintentionally influenced by a student's reading ability as opposed to their knowledge of addition and subtraction.

Sources of Validity Evidence

The Standards for Educational and Psychology Testing (2014) is a joint publication of the American Educational Research Association, American Psychology Association, and the National Council of Measurement Education. This text provides criteria that can be used to evaluate the appropriateness of a test. The Standards describe the following sources of evidence: (1) content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) test consequences. Each source of evidence is briefly reviewed in turn.

Content

Test content should be *relevant* and *representative*. Irrelevant content, or content that fails to be indicative of the aim of a measurement procedure, should obviously be avoided. For example, an individual who is interested in measuring self-esteem may need to clearly differentiate this concept from depression. A failure to differentiate these concepts may lead to inadvertently confounding indicators of depression with self-esteem.

Test content should also represent the full breadth of what one is attempting to measure. In other words, it is important that the content reflects each aspect of a given attribute or construct. For example, if depression is theoretically viewed as having cognitive, behavioral, and affective manifestations then it is important that a sufficient number of items are used to indicate each of these aspects of depression.

A common way to view test content is through a specification table. Basically, this table indicates the number or proportion of items on a test aiming to assess a given attribute or skill. In an achievement test experts may believe that "objective 1" should be weighed twice as heavily as "objective 2." If you have a 30 item test, 20 items would aim to measure objective 1 and 10 items would aim to measure objective 2. Content experts are frequently used as a means to examine the relevance and representativeness of a set of items.

Response Processes

Response processes generally refer to the cognitive processes used by a respondent to answer an item or a set of items. This is perhaps one of the most challenging sources of evidence to accumulate. However, accumulating this type of evidence is critical whenever score-based interpretations emphasize such processes. For example, for a math test our interpretation of scores may emphasize the use of particular strategies to answer a set of items. Children with more complex mathematical reasoning may rely upon some strategies more so than others. In this case, it may be possible to have children show their work when solving problems to investigate solution strategies.

In other cases it may be possible to design "distractors," or answers that are incorrect on a multiple choice exam, to align with specific errors in thinking. Selection of these distractors may indicate common misconceptions that are important to rectify in subsequent instruction. A different line of evidence may consist of cognitive interviewing to investigate whether respondents are interpreting an item in an intended way. Issues would exist if a substantial number of respondents indicate unintended interpretations.

Internal Structure

An examination of internal structure calls for investigating the extent to which relationships among items conform to the attribute one is attempting to measure. Such an investigation generally, asks: "How many do I seem to be measuring?" Consider an example in which items are written to reflect cognitive, behavioral, and affective aspects of depression. Each of these aspects may be viewed as separate but related elements of depression. In this case we should expect to see that items referring to behavioral aspects of depression are more correlated with each other than with items referring to a different aspect of depression.

Various procedures may be used to investigate whether the pattern of correlations/covariances make theoretical sense. Investigations of internal structure typically include exploratory factor analysis, confirmatory factor analysis, and/or item response theory.

It should also be mentioned that an examination of internal structure is also concerned with differential item functioning. *Differential item functioning* exists when people of the same ability (or level of an attribute) have a different probability of getting an item correct (or endorsing an item). For example, if males and females of the same ability have a different probability of getting an item correct then this would be a validity issue pertaining to the internal structure of the test (see psychometric theory texts at the end of the guidebook for more information about this topic).

Relations with Other Variables

This source of evidence includes various considerations and/or types of studies. For example, one may investigate convergent and discriminant evidence. In other words, if a measure is an indication of reading ability then it should correlate with other similar measures (*convergent*). However, it should be unrelated to other variables such as administration format (*discriminant*). Hypothesized group-differences and experimental evidence pertaining to score-based inferences is also included in this category.

Finally, this source is also concerned with test-criterion relationships. A criterion is basically something that would like to predict from a measure. For example, SAT/ACT scores are primarily used to predict college grade point average. Other measures may be concerned with predicting retention or job readiness. In such cases, one may examine the extent to which a measure predicts the criterion. Criterion variables may be measured at the same time as your instrument (i.e. *concurrent*) or at some point in the future (i.e. *predictive*).

Consequences of Testing

Consequences of testing may be categorized as intended or unintended. A potential user of a test may argue that an instrument should be adopted because of positive consequences. It may be suggested, for example, that a reason to adopt a performance assessment (e.g. constructed response) is because it has a positive impact on pedagogy within the classroom. If such an argument is made, then it is important to examine the extent to which these consequences are realized.

Unintended consequences may occur in numerous ways. For example, administering a test may inadvertently lead to differential selection of specific groups into an educational program. Such consequences are a validity issue if they stem from one of the sources of invalidity previously discussed (i.e. construct underrepresentation or construct-irrelevant variance). Unintended consequences may also result from reasons that fail to be a validity issue. For example, it is possible that administering a test creates a hostile environment in a school setting or leads to decrease teacher motivation. In such situations, an individual must weigh the costs and benefits of testing to determine whether a given instrument is suitable for their purpose.

GENERAL RECOMMENDATIONS: PUTTING THE PUZZLE TOGETHER

This handbook provided an overview of the concepts of reliability and validity with respect to instrument selection and design. This overview described direct versus indirect measures of student learning and/or development, verb-instrument agreement, and outcome-instrument maps. While "directness" is more of a continuum than a dichotomy, one should determine the type of information that is of interest from students (e.g., self-report or content knowledge) in order to select the most appropriate instrument. Instruments should be selected that are more direct, given a stated aim of inquiry. The student learning outcomes should also 1) clearly align with the assessment instruments selected and 2) use clear verbs that serve as a hint about what type of assessment strategy is applicable.

There are advantages and disadvantages to either selecting an existing instrument or designing a new instrument. While preexisting measures are more convenient, strong reliability and validity evidence are necessary to justify any interpretations or uses of the test scores. Just because the name of an instrument seems to match your objectives does not mean it is a good measure. With this said, developing a new instrument can be time consuming and expensive, but has an advantage of likely being more aligned with your objectives.

We have considered various strategies for writing cognitive and attitudinal items. Matching the type of item to the desired learning outcomes is a first step in this process. Closed-ended items, such as multiple choice, true/false, and matching items are relatively easy to score. However, these types of items are also susceptible to guessing. Open-ended items, such as short answers and sentence completions are more difficult to score. This technique usually requires one to develop a checklist or rubric for scoring purposes. Attitudinal items do not have right or wrong answers, so guessing is not a concern. However, other issues can inadvertently influence how an individual responds to these items, such as social desirability and the use of leading statements. It should be clear that no single item type is "better" than the other. Rather, different item types are appropriate for different types of learning outcomes and cognitive levels.

Reliability is broadly concerned with the consistency of scores on an instrument, while error is an indication of score inconsistency. Scores can be inconsistent over time, across different forms (or versions) of a test, or across a set of items written to measure the same attribute. Since reliability is a property of the scores (not the test itself), you should always estimate reliability on your sample. This also implies that it can be dangerous to assume that scores in your study will have a similar level of reliability as what is published in prior research.

According to most theorists, validity is property of inferences as opposed to being a property of a test. Strictly speaking, tests are not valid or invalid. Instead interpretations of scores for proposed uses of a test are more or less valid. Validation is a process used to investigate interpretations and uses of a test. The Standards for Educational and Psychological Testing (2014) address five sources of validation evidence which includes content, response processes, internal structure, relations to other variables, and test consequences. Evidence should also be collected in an effort to rule out primary threats to validity (i.e. construct underrepresentation and construct-irrelevant variance).

In conclusion, this guidebook has aimed to provide a general introduction to measurement issues when selecting and/or designing an instrument. Various topics were excluded and there have been entire books written about the topics we have chosen to include. The information provided is not by any means exhaustive, though we are hopeful that it serves as a valuable resource to those who are interested in obtaining a basic overview of measurement issues. Additional resources are provided below for those who are interested in extending their knowledge about this important topic.

ADDITIONAL RESOURCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological tests.* Washington, DC: American Educational Research Association.

DeMars, C. (2010). *Item response theory*. New York, NY: Oxford University Press.

DeVellis, R.F. (2012). *Scale development: Theory and applications* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Furr, R.M. & Bacharach, V.R. (2014). *Psychometrics: An introduction* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Haladyna, T.M. & Rodriguez, M.C. (2013). *Developing and validating test items.* New York, NY: Routledge.

Hathcoat, J.D. (2013). Validity semantics in educational and psychological assessment. *Practical Assessment, Research, and Evaluation*, *18*, 1-14.

Hathcoat, J.D. (2013). Validity semantics in educational and psychological assessment.

Hathcoat, J.D. (2015). Introduction to validity theory: Validity 101 [Video file]. Retrieved from https://www.youtube.com/watch?v=rYc-coraFNk

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational and Psychological Measurement*, *38*, 319-424.

Krathwohl, D.R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41(4),* 212-218.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington DC: American Psychological Association.