Psychometrics and Scale Construction

James Jaccard
New York University

Overview

Reliability and validity

Random and systematic measurement error

Latent variable models of measurement

Item level models of measurement/factor analysis

Concept mapping

Writing items and item metrics

Designing a psychometric study

Overview

Coefficient alpha and some alternatives

Evaluating extant measures

Best practices

Measurement and Science

Measurement and Science

Measurement is key to science. Without measurement, there can be no science

There are many "systems of knowing" (philosophy, religion, art, literature). The defining characteristic of science is that knowledge must be grounded in empirical verification

Measurement is key to empirical verification and most any concept can be measured!

Measurement and Science

Good research gets its house in order measurement-wise before major data collection commences

Large, secondary data bases often are problematic because they tend to sacrifice measurement quality for concept breadth.

It is essential in such research to use "measurement forgiving" analytic methods (SEM).

Reliability and Validity

Validity

When individuals complete a test or a scale, their responses, in theory, are impacted by their standing on the underlying construct of interest

The extent to which this is true (and if the measure is impacted by nothing but the underlying construct), the measure is said to be *valid* or that it "*has validity*."

Random Error

Sometimes a person's response is impacted by factors other than the underlying construct

One source of "noise" on a measure is called random error. This reflects random influences that arbitrarily push scores up or down:

Distractions

Misreading an item

Item ambiguity

Reliability

Reliability is the extent to which a measure is *free of random error*.

If the reliability of a measure is 0.90, then 90% of its variation is systematic and 10% of its variation is random "noise"

If the reliability of a measure is 0.80, then 80% of its variation is systematic and 20% of its variation is random "noise"

If the reliability of a measure is 0.70, then 70% of its variation is systematic and 30% of its variation is random "noise"

Reliability

Reliability is not the extent to which scores on a scale at one point in time correlate with scores at another point in time

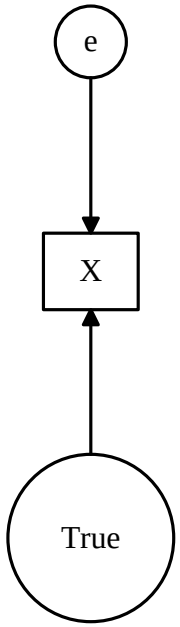
Reliability is not coefficient alpha

These are (imperfect) methodological strategies for *estimating* the reliability of a measure. They are not reliability *per se*

Latent Variable Representations of Measurement

Measurement Model with Random Measurement Error

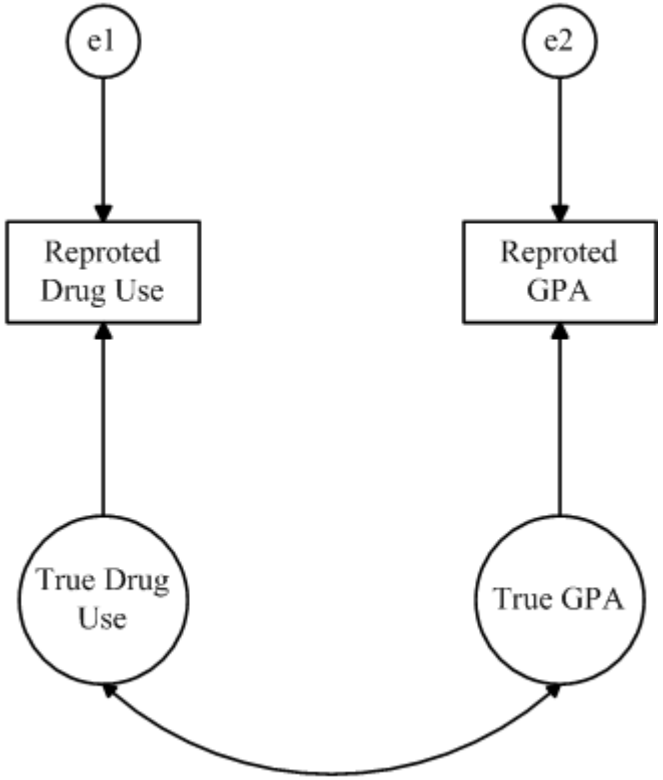
(a)



(b)

Adverse Effects of Random Measurement Error

Random Measurement Error



Random measurement error attenuates the observed correlation between variables

(if two measures each have reliability of 0.70, and the true correlation is 0.50, the measured correlation will be 0.35)

Random Measurement Error

In multiple regression, random measurement error undermines control of covariates

$$GPA = a + b_1 \text{ Depression } + \ b_2 \text{ Income } \ + \ e$$

(You think you have controlled for income, but you really have not because your measure of income is not a pure measure of income – it has random noise in it)

Reducing Unreliability

Reducing Unreliability

Practices to minimize unreliability:

1. Use an accommodating test environment – heat, lighting, noise, presence of others

2. Minimize respondents rushing through the task by keeping assessment burden low; split measures into packets

3. Give respondents practice items so they accommodate to the task, the rating scale, and the testing environment

Reducing Unreliability

4. Ensure instructions are clear; confusing directions lead to confused respondents and random error

5. Ensure your items have no ambiguities

"I have smoked marijuana in the past month"

"I have smoked marijuana in the past 30 days"

6. Use multiple items to assess your construct

The Strategy of Using Multiple Items

If we assess a construct using multiple items, when we average across items, random errors cancel, yielding a more reliable index:

	Observed Score	True Score	Error Score
Item 1	4	3	+1
Item 2	3	3	0
Item 3	2	3	-1
Average	3	3	0

The more items, the better the cancelation process works

The Strategy of Using Multiple Items

Some constructs are so straightforward and unambiguous that unreliability is of little concern (and use of multiple items is silly)

What is your age? ____ years old

What is your gender? ____ Male ____ Female

The Strategy of Using Multiple Items

Trade-offs of using multiple items if the time you have to collect data is limited:

- using multiple items increases reliability, but….

- using multiple items for a single construct decreases the number of constructs you can assess

Need to balance breadth of assessments with quality of measures

Psychometric Best Practices

Psychometric Best Practices

Practice 1: Adopt the six practices to minimize unreliability

Structural Equation Modeling

When reliability of measures falls below 0.70, this is seen as problematic because such levels of random noise typically play havoc on conclusions in complex multivariate modeling

SEM analytic approaches can provide direct perspectives on how random measurement error affects conclusions and can make adjustments for its presence (even with single items)

SEM is preferable to multiple regression when measures have random measurement error (if random error is minimal, multiple regression is viable)

Psychometric Best Practices

Practice 1: Adopt the six practices to minimize unreliability

Practice 2: Use SEM to gain perspectives on the impact of unreliability of measures on conclusions

Systematic Measurement Error

Systematic Error

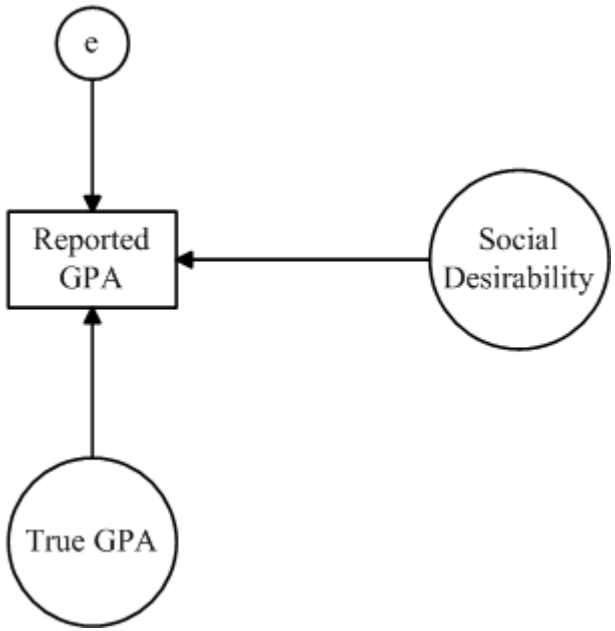
Another type of error that afflicts measures is systematic error, which are non-random factors that impact scores

Constant error

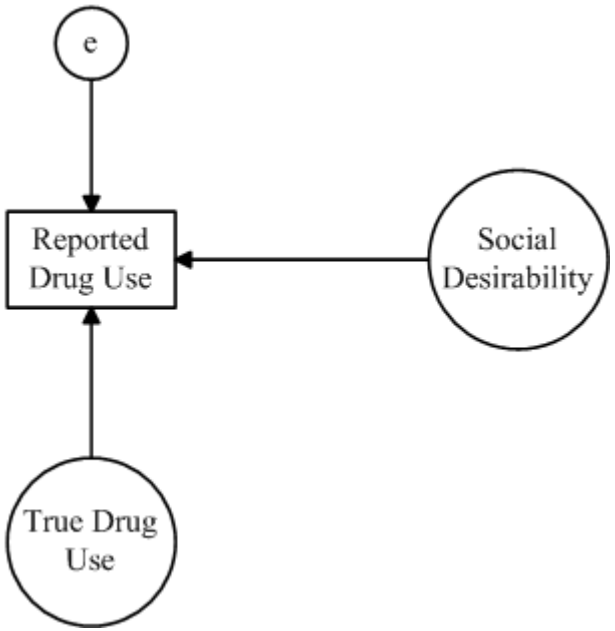
Good impression/social desirability

Model with Random/Systematic Measurement Error

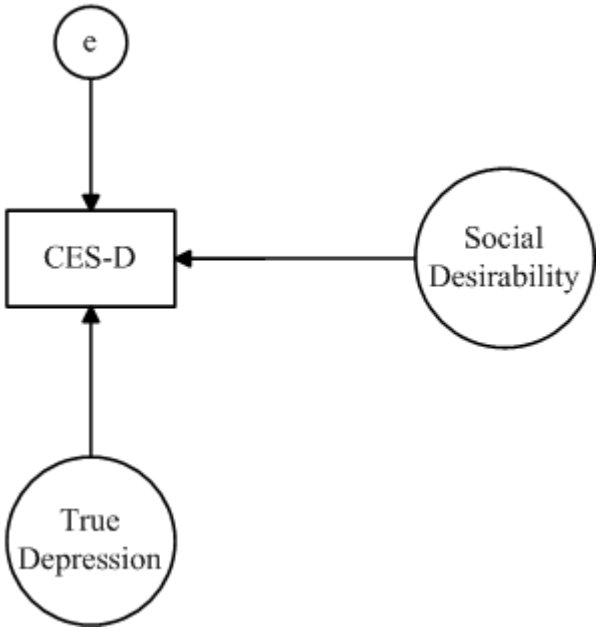
(a)



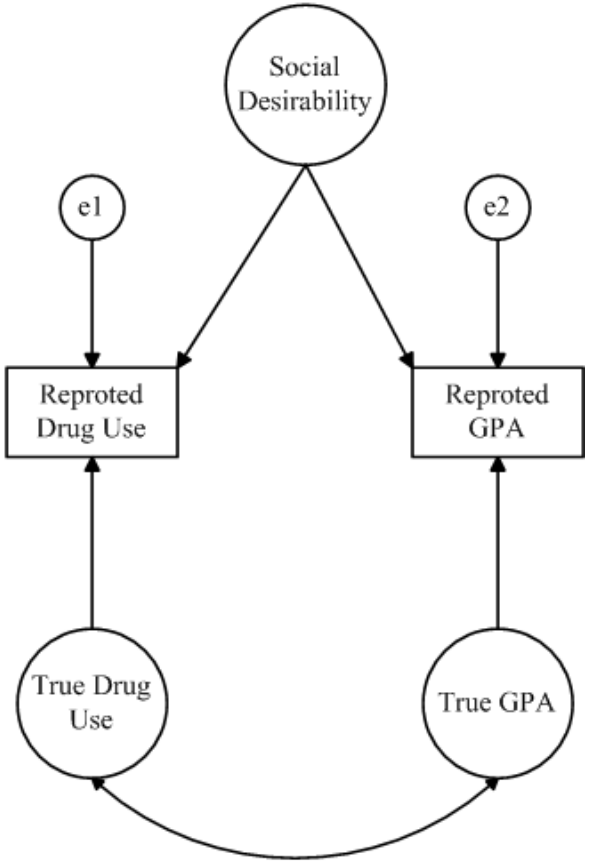
(b)



Model with Random/Systematic Measurement Error



Systematic Measurement Error



Systematic measurement error can inflate the observed correlation between variables

Systematic Measurement Error

In multiple regression, systematic measurement error undermines control of covariates and biases estimates of coefficients

$$GPA = a + b_1\ Depression\ +\ b_2\ Income\ +\ e$$

(All of the correlations between these variables are inflated by the common influence of social desirability on them. This produces biased coefficients and inaccurate p values and standard errors)

Reducing Good Impression Influences

1. Use self administered as opposed to face-to-face reports (use of mixed strategies; CASI)

2. Use anonymous or confidential conditions and ensure respondents know this

3. Provide motivational instructions to be honest

4. Instruct to skip answering if can't be truthful

5. Obtain a measure of social desirability and use it as a statistical covariate in your modeling

Social Desirability/Good Impressions

The most frequently used SD scale is the Crowne-Marlowe scale, which has poor psychometric properties[1]

None of the extant measures do a good job at measuring the type of defensiveness we seek to correct in survey work

Instead, they measure a general trait-like quality subsuming need for approval, vulnerable self esteem, and repressiveness, especially in interpersonal situations[1,2]

For detailed reviews, see Uziel (2010)[1] and Fleming (2012)[2]

Social Desirability/Good Impressions

Most SD scales are too long and not practical

I use a 4 item index with disagree-agree metrics:

> **I never swear; I never say something bad about a friend behind his/her back; I never criticize other people; I never get sad**

Items are adapted from the BDRI impression management subscale of Paulhus, which is one of the better measures[3]

Psychometric Best Practices

Psychometric Best Practices

Practice 1: Adopt the six practices to minimize unreliability

Practice 2: Use SEM to gain perspectives on the impact of unreliability of measures on conclusions

Practice 3: Adopt the five practices to minimize social desirability/good impression influences on measures

Structural Equation Modeling

SEM analytic approaches can provide direct perspectives on how systematic measurement error affects conclusions and can make adjustments for its presence

SEM is preferable to multiple regression when measures have systematic measurement error (if systematic error is minimal, multiple regression is viable)

Psychometric Best Practices

Practice 1: Adopt the six practices to minimize unreliability

Practice 2: Use SEM to gain perspectives on the impact of unreliability of measures on conclusions

Practice 3: Adopt the five practices to minimize social desirability/good impression influences on measures

Practice 4: Use SEM to gain perspectives on the impact of systematic measurement error on conclusions

More on Systematic Error

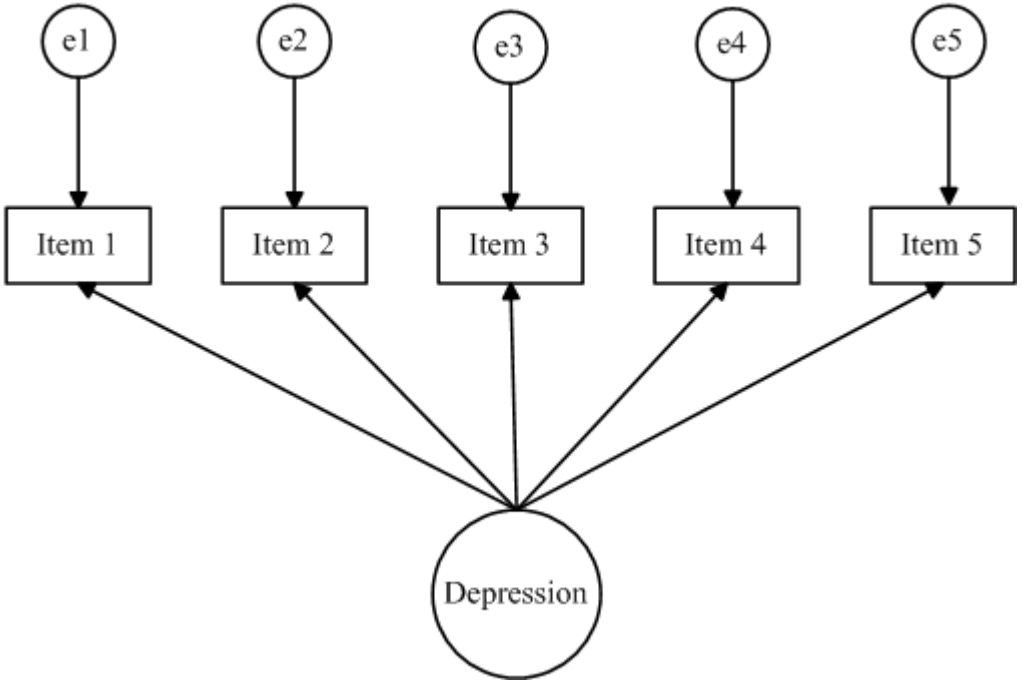
Psychometricians have identified other response styles, including (a) an acquiescence response set, (b) a disacquiescence response set, and (c) a middle category response set

The empirical evidence for the importance of these artifacts is inconsistent and, overall, research suggests their effects are weak

For (conflicting) reviews of response set literatures, see Conway and Lance (2010),[4] Podskoff et al. (2012),[5] Borer (1965),[6] and Wiggins (1973).[7]

Latent Variable Measurement Models at the Item Level

Item Level Model



With no random error, items should be highly correlated. The more random error, the lower the item correlations

Item Level Model

If items are correlated 0.20, there is about 80% random noise in each item

If items are correlated 0.30, there is about 70% random noise in each item

If items are correlated 0.40, there is about 60% random noise in each item

(What is a reasonable degree of random error to have operate to still qualify as a "good" item?)

Item Level Model

If we average items into a total score (so item errors cancel), we can tolerate more item-level noise (entries are reliabilities)

Number of Items	Proportion of Item Level Noise									
	0.65	0.60	0.55	0.50	0.45	0.40	0.35	0.30	0.25	0.20
2	0.51	0.57	0.62	0.66	0.71	0.75	0.78	0.82	0.85	0.88
3	0.61	0.66	0.71	0.75	0.78	0.81	0.84	0.87	0.90	0.92
4	0.68	0.72	0.76	0.80	0.83	0.85	0.88	0.90	0.92	0.94
5	0.72	0.76	0.80	0.83	0.85	0.88	0.90	0.92	0.93	0.95
6	0.76	0.80	0.83	0.85	0.88	0.90	0.91	0.93	0.94	0.96
7	0.79	0.82	0.85	0.87	0.89	0.91	0.92	0.94	0.95	0.96
8	0.81	0.84	0.86	0.88	0.90	0.92	0.93	0.95	0.96	0.97
9	0.82	0.85	0.88	0.90	0.91	0.93	0.94	0.96	0.96	0.97
10	0.843	0.87	0.89	0.91	0.92	0.94	0.95	0.96	0.97	0.98

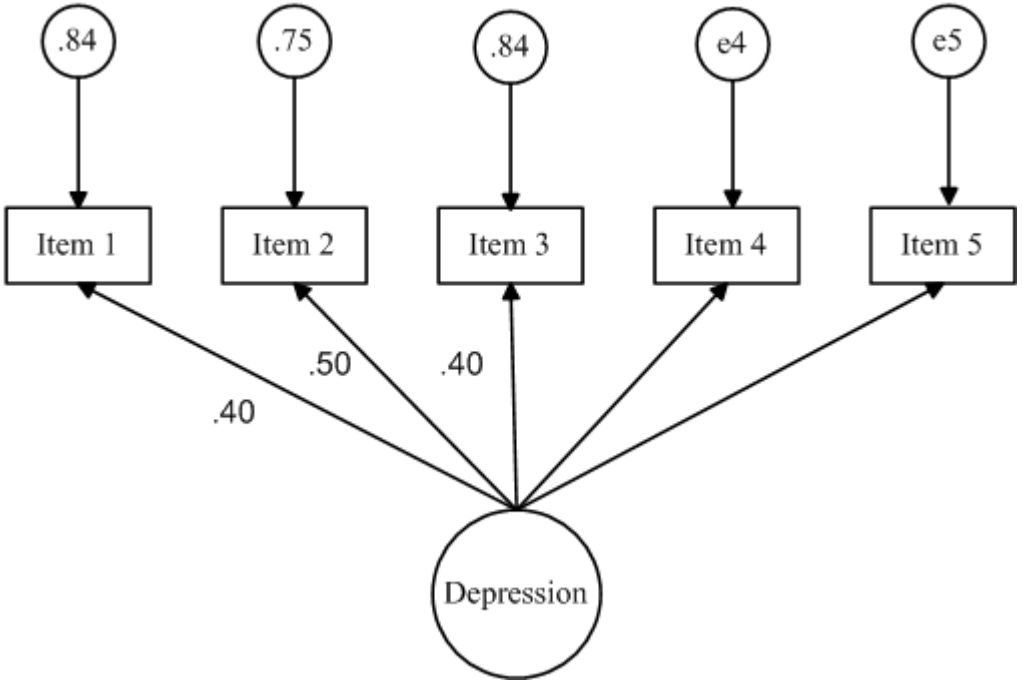
Item Level Model

Here is the same table but expressed using average item inter-correlations

Number	Item Level Correlations									
Items	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80
2	0.51	0.57	0.62	0.66	0.71	0.75	0.78	0.82	0.85	0.88
3	0.61	0.66	0.71	0.75	0.78	0.81	0.84	0.87	0.90	0.92
4	0.68	0.72	0.76	0.80	0.83	0.85	0.88	0.90	0.92	0.94
5	0.72	0.76	0.80	0.83	0.85	0.88	0.90	0.92	0.93	0.95
6	0.76	0.80	0.83	0.85	0.88	0.90	0.91	0.93	0.94	0.96
7	0.79	0.82	0.85	0.87	0.89	0.91	0.92	0.94	0.95	0.96
8	0.81	0.84	0.86	0.88	0.90	0.92	0.93	0.95	0.96	0.97
9	0.82	0.85	0.88	0.90	0.91	0.93	0.94	0.96	0.96	0.97
10	0.843	0.87	0.89	0.91	0.92	0.94	0.95	0.96	0.97	0.98

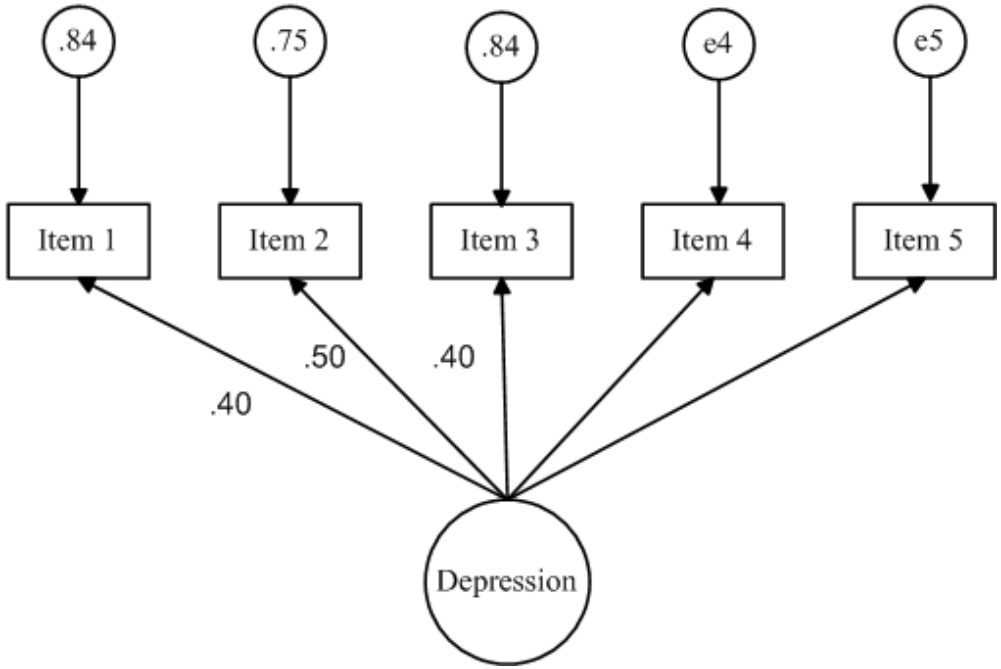
A Guide to Factor Analyses of Items

Factor Analyses of Items



One minus the squared factor loading is the proportion of variation in the item that is random noise

Factor Analyses of Items



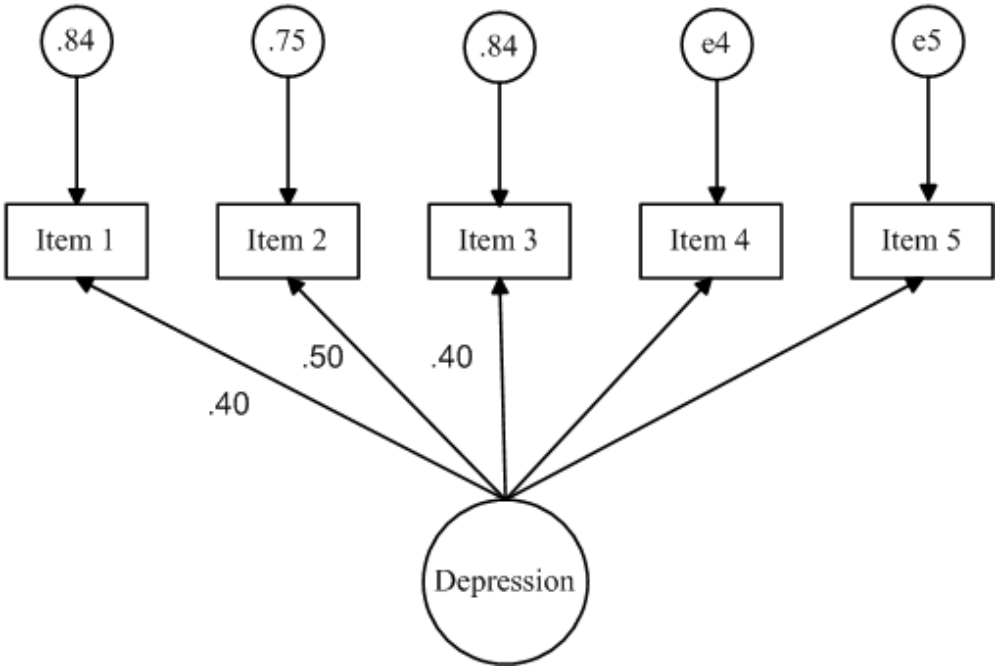
Item 1: It is hard for me to get going in the morning

Item 2: I often feel sad

Item 3: I sometimes think life is not worth living

Instead of conceptualizing errors as random noise, we often think of them as variance unique to the item that is due to factors other than the construct

Factor Analyses of Items



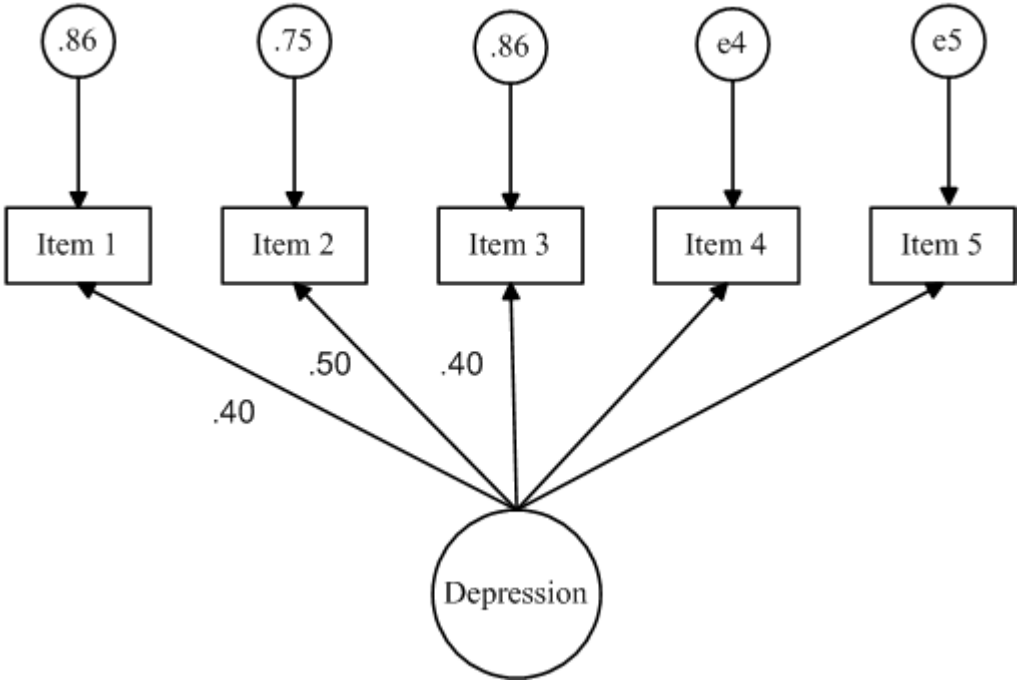
Item 1: It is hard for me to get going in the morning

Item 2: I often feel sad

Item 3: I sometimes think life is not worth living

Unique variance is measurement error; but we label it and think of it a bit differently than random "noise"

Factor Analyses of Items



The product of two factor loadings for items reflecting the same factor equals the correlation between the items

Stigma Discrimination due to Mental Health

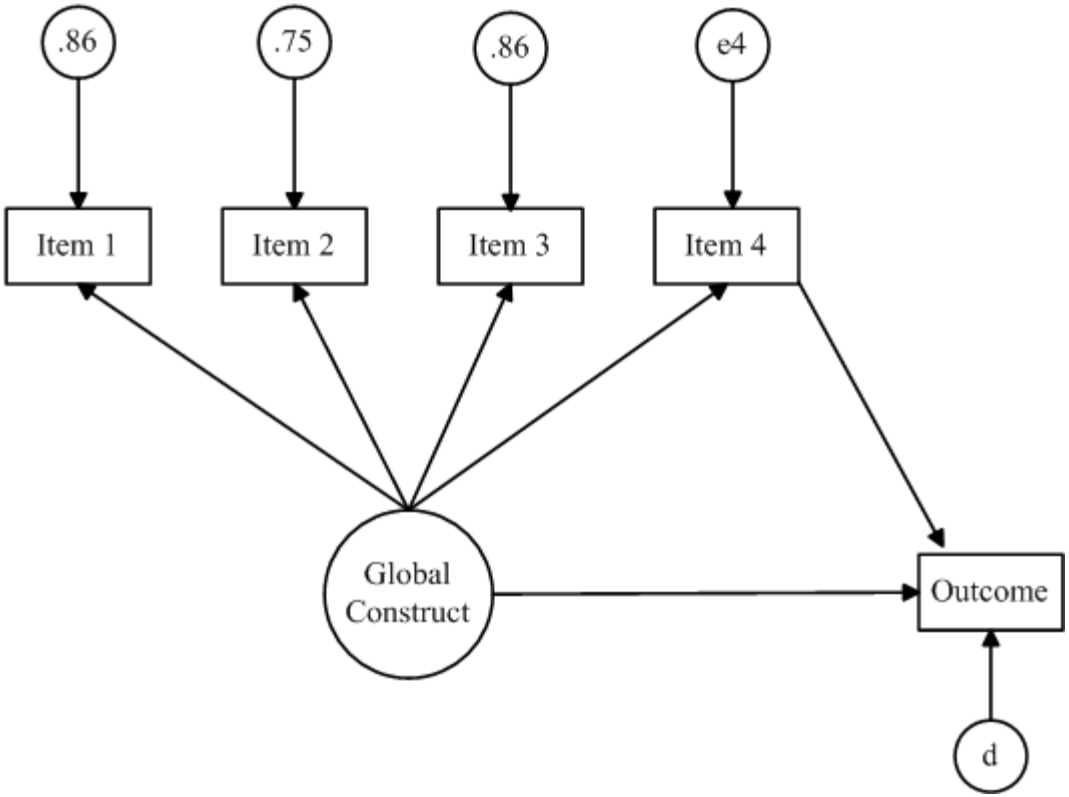
	Loading
Sometimes I am talked down to because of my mental health problems	.52
I have been discriminated against by employers because of my mental health problems	.43
I would have had better chances in life if I had not had a mental illness	.61
People's reactions to my mental health problems make me keep to myself	.49

Stigma Discrimination due to Mental Health

We are quick to focus on common global variance by collapsing (averaging) across items that have a great deal of unique variance.

Perhaps we are “shooting ourselves in the foot” theoretically by ignoring such substantial portions of unique variation

Stigma Discrimination due to Mental Health



The Importance of Theory: Mapping the Concept

Mapping the Concept

Some constructs are narrow in scope and specific (e.g., gender, age, annual income)

Other concepts (e.g., intelligence , depression, and social support) are abstract, complex, and multidimensional

To measure a construct, you must clearly define it and then specify the multidimensional structure of it on conceptual grounds (or through qualitative work).

Mapping the Concept

Depression

Cognitive, affective, somatic

Social Support

Tangible, emotional, informational

Anxiety

Social, generalized, panic

Mapping the Concept

School connectedness is the extent to which students feel personally accepted, respected, included, and supported by others in the school social environment (Goodenow)

***Weak approach to mapping*: Generate items without thinking through if there are different facets (framed by theory, past research, or qualitative work) that need to be represented**

***Strong approach to mapping*: Map fundamental facets that need to be measured to comprehensively represent the construct (framed by theory, past research, or qualitative work). Generate items for each facet.**

Mapping the Concept

Treat each facet as a concept in its own right that you will need to generate items to assess

For extant scales, read the original articles that developed the scale. Examine the theoretical bases of concept mapping.

Was a weak or strong approach used?

(Exceptions are purely empirical approaches, like the big five personality traits and analyses of the structure of emotions)

Psychometric Best Practices

Practice 1: Adopt the six practices to minimize unreliability

Practice 2: Use SEM to gain perspectives on the impact of unreliability of measures on conclusions

Practice 3: Adopt the five practices to minimize social desirability/good impression influences on measures

Practice 4: Use SEM to gain perspectives on the impact of systematic measurement error on conclusions

Practice 5: Use strong rather than weak approaches to concept mapping

Ten Tips for Writing Items

Ten Tips for Writing Items

Good items have the following properties:

1. They are highly correlated with the underlying construct

Hillary Clinton is a Democrat

As you write an item, ask yourself: "will people 'high' on the construct respond differently to this item than people who are 'low' on the construct"

Hillary Clinton would be a good leader

Ten Tips for Writing Items

2. They do not have base rate problems (floor and ceiling effects)

Hillary Clinton is a Democrat

For me, getting pregnant now would be bad

You can alter response distributions to an item by making it more or less extreme

For me, getting pregnant now would be one of the worst things that could happen to me

Last week I was sad **versus** ***Last week, I was very sad***

Ten Tips for Writing Items

3. They are short, simple, and understandable

You can apply readability formulae available in Microsoft Word or commercial software [Flesch-Kincaid; Simple Measure of Gobbledygook (SMOG)][8-9]

No big words and long sentences!

Using Reading Formulae

Sometimes I am talked down to because of my mental health problems.

I have been discriminated against by employers because of my mental health problems.

I would have had better chances in life if I had not had a mental illness.

People's reactions to my mental health problems make me keep to myself.

Ten Tips for Writing Items

4. They have a single thought (are not double barreled)

My therapist was expert and sincere

I intend to go to my appointment because it will help me get better

5. They do not contain double negatives

I should not oppose the Affordable Care Act

Ten Tips for Writing Items

6. They have no ambiguities

I have used drugs in the past year

7. They personalize and provide contextual and time frames

For me, joining a gang in my neighborhood at this time would be good

8. They avoid abbreviations and slang

I know the whereabouts of my child 24/7

Ten Tips for Writing Items

9. They are not leading

My mother disapproves of me smoking marijuana at this time in my life

(conflicts with advice to minimize metric shifts)

Ten Tips for Writing Items

10. They are phrased to minimize good impression tendencies (e.g., some research suggests adults are more comfortable reporting the year they were born rather than their age; use of age categories rather than age per se)

Use reminder instructions for honest answers or give an "out" for a non-socially desirable response (*There are many reasons why people don't get a chance to vote. Sometimes they have an emergency, or are ill, or simply can't get to the polls. Did you vote in the last election?*)

Writing Items

Generate more items than you want in your scale, as you invariably will eliminate badly behaved items during the scale construction process.

Try to keep your scale brief and practical

The items for a scale (or subscale) should be thought of as homogenous and interchangeable – all reflecting the same thing

Psychometric Best Practices

Practice 6: Use the 10 strategies for writing items

Item Metrics

Item Metrics

Items are typically rated on a scale metric:

True-False

Agree-Disagree

Approve-Disapprove

Favorable-Unfavorable

The metric can be dichotomous or many-valued: strongly agree, moderately agree, neither, moderately disagree, strongly agree

Item Metrics

The precision of a metric or scale is the number of discriminations it makes

Scales vary widely in the precision of the metrics used

Lack of precision is problematic because it places people with meaningfully different opinions into the same category

Do you approve or disapprove of the Affordable Care Act?

___ Disapprove ___ Approve

Item Metrics

As a general rule, you want to avoid dichotomous metrics and use metrics that make at least 4, preferably 5 (or more) discriminations

(simulation work by Bollen[10] and others[11])

For populations with low education or for whom rating metrics are not viable, you can use two step approaches to maintain precision (e.g., Goldin's research in Guatemala)

Adverb Qualifiers

Research suggests it is better to use adverb qualifiers than not[12]

Strongly Disagree	0	1	2	3	4	5	Strongly Agree

___ Strongly agree
___ Moderately agree
___ Neither
___ Moderately disagree
___ Strongly disagree

Adverb Qualifiers for Magnitude Judgments

Different adverbs have different modifying values.

- **"Slightly" good is half (0.50) as good as the unmodified good**
- **"Extremely" good is 1.5 times as good as the unmodified good**

There are published lists of the modifying values of different adverbs (see Beckstead for a review and examples[12])

You should select adverbs that are "equally spaced" (or that produce equal intervals) across the dimension

Adverb Qualifiers

Do not take the lists too literally as the values vary from population to population and there is sampling error in them

Similar psychometric work has been conducted for expressions of frequency (never, occasionally, often, always)[12-13] and agree-disagree scales[14]

Agreement Examples

Strongly Agree
Moderately Agree
Neither
Moderately Disagree
Strongly Disagree

Strongly Agree
Agree
Neither
Disagree
Strongly Disagree

Frequency Examples

Very Frequently	Always or almost always
Frequently	Usually
Occasionally	About Half the Time
Rarely	Sometimes
Never	Never or almost never

Importance and Bipolar Examples

Extremely favorable
Quite favorable
Slightly favorable
Neither
Slightly unfavorable
Quite unfavorable
Extremely unfavorable

Extremely important
Quite important
Slightly important
Not at all important

Very Important
Moderately Important
Slightly important
Unimportant

Mixed Approach

Sometimes to maximize precision, we combine the use of numbers with labels

0	1	2	3	4	5	6	7	8	9	10
Not at all certain			slightly certain			quite certain			extremely certain	

Multi-Item Scales

If we use dichotomous item metrics, with multi-item scales we still obtain reasonable precision when we sum or average across items.

But, we make item-level analyses (e.g., test of unidimensionality) more challenging

Studies have found that using more precise metrics at the item level for multi-item tests often improves reliability and validity

Psychometric Best Practices

Practice 6: Use the 10 strategies for writing items

Practice 7: Address issues of metric precision

Practice 8: Select adverb qualifiers and phrases that maximize interval level properties of the metric

Cognitive Testing

Cognitive Testing

To improve the quality of items, we often conduct cognitive testing

This involves having 3 to 5 individuals from the target population complete the survey and provide reactions to it.

One approach uses think aloud strategies in which people verbalize whatever they are thinking as the read and respond to each question

Cognitive Testing

Disadvantages of think aloud strategy:

- **Respondents go off on tangents**
- **Respondents focus more on response process than on questions**
- **Process of thinking aloud may change answering process**
- **Respondents don’t necessarily provide all types of useful information**

Cognitive Testing

Another approach uses probe strategies, soliciting feedback after person completes each question or a subset of questions

Issues of ambiguity, ease of responding, comprehension, readability, honesty, and interest value are probed

- **"What made you say that?"**
- **"Why did you respond that way?"**
- **"What does that mean to you?"**
- **"Tell me what I was asking in your own words?"**

Cognitive Testing

- "Was anything ambiguous?"
- "Was there anything or words you did not understand?"
- "Level with me – did you respond honestly?"
- "Did you find the question easy to answer? Why not?
- "Was this interesting or boring?

Items are refined and then subjected to more formal empirical evaluation

Willis,G. (2004). Cognitive interviewing: A tool for improving questionnaire design. Newbury Park: Sage

Psychometric Best Practices

Practice 6: Use the 10 strategies for writing items

Practice 7: Address issues of metric precision

Practice 8: Select adverb qualifiers and phrases that maximize interval level properties of the metric

Practice 9: Apply cognitive testing to refine items

Anchoring

Anchoring

It also is good practice to anchor the extremes of the scale, giving respondents examples of responses on the low and high end of the metrics (i.e., ground the metric)

Anchoring is commonly used in organizational studies of performance (e.g., ratings of employee effectiveness) in which one provides examples of ineffective and effective behavior to "anchor" scale extremes and midpoint (BARS methodology).[15]

Studies of caseworker ratings of incidence severity are impacted by exemplars used to anchor the scale endpoints[16]

Psychometric Best Practices

Practice 6: Use the 10 strategies for writing items

Practice 7: Address issues of metric precision

Practice 8: Select adverb qualifiers and phrases that maximize interval level properties of the metric

Practice 9: Apply cognitive testing to refine items

Practice 10: Use anchoring

A Psychometric Study

A Psychometric Study

We next conduct a psychometric study to guide final scale development and measure evaluation

We want to power the study to detect medium sized correlations and to yield margins of error for reliability estimates that are acceptable

We also want to have a sample size that is sufficient for confirmatory factor analysis in terms of asymptotic theory and stable covariance matrices

A Psychometric Study

We want to address issues of reliability. Administer survey in a test-retest format with a two week interval between assessments

We want to address issues of discriminant validity. Administer an index of social desirability response tendency

We want to address issues of construct validity. Include measures of constructs that permit this.

Data Analysis

Eliminate any item with floor or ceiling effects (more than 90% are in one response category

Eliminate any item with poor test-retest reliability correlation (less than 0.50 for multi-item scales)

Eliminate any item that shows a sizeable correlation with social desirability scales (correlation of 0.33 or greater)

As you eliminate items, make sure you remain true to your conceptual map

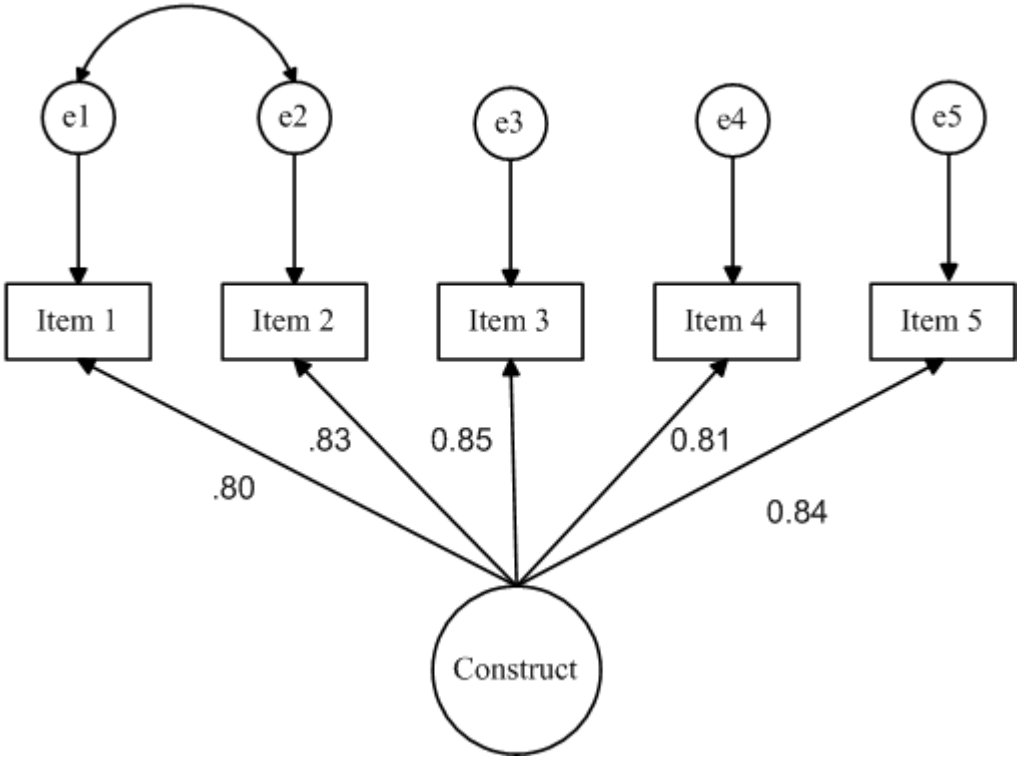
Data Analysis

Perform a confirmatory factor analysis in accord with the conceptual map of the construct

Identify and eliminate poorly behaved items in terms of modification indices for large correlated errors (.30 or higher) and low factor loadings (less than 0.50)

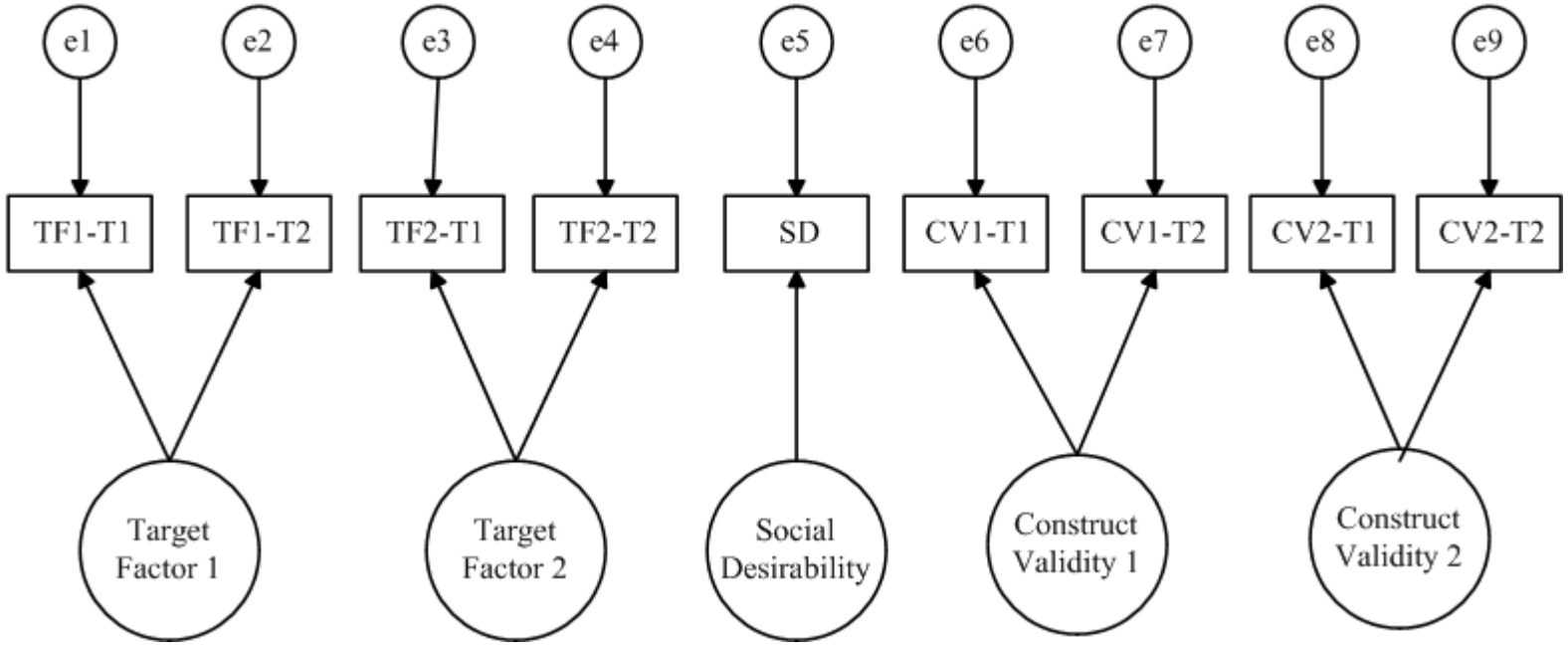
Consider a dominant single factor solution versus a multi-factor solution for the total scale as well as each subscale.

Data Analysis



Data Analysis

To estimate reliability and validity, calculate total scores for each construct and then fit the CFA model:



(All latent variables are correlated)

Psychometric Best Practices

Practice 11: Conduct a psychometric study to demonstrate construct mapping, reliability and validity

The Limitations of Coefficient Alpha

The Limitations of Coefficient Alpha

PSYCHOMETRIKA—VOL. 74, NO. 1, 107–120
MARCH 2009
DOI: 10.1007/S11336-008-9101-0

ON THE USE, THE MISUSE, AND THE VERY LIMITED USEFULNESS OF CRONBACH'S ALPHA

KLAAS SIJTSMA

TILBURG UNIVERSITY

This discussion paper argues that both the use of Cronbach's alpha as a reliability estimate and as a measure of internal consistency suffer from major problems. First, alpha always has a value, which cannot be equal to the test score's reliability given the interitem covariance matrix and the usual assumptions about measurement error. Second, in practice, alpha is used more often as a measure of the test's internal consistency than as an estimate of reliability. However, it can be shown easily that alpha is unrelated to the internal structure of the test. It is further discussed that statistics based on a single test administration do not convey much information about the accuracy of individuals' test performance. The paper ends with a list of conclusions about the usefulness of alpha.

Key words: Cronbach's alpha, internal consistency, reliability, unidimensionality.

Limitations of Coefficient Alpha

Coefficient alpha is not a good test of unidmensionality. CFA is better

Coefficient alpha is an inferior indicator of reliability. It makes too many assumptions (equal factor loadings and uncorrelated measurement errors)

Alternative index is the composite reliability (CR), which is calculated in the context of CFA[17]

$CR = (\Sigma p)^2 / [(\Sigma p)^2 + (\Sigma e)]$, where p = the factor loadings and e = the error variances

Psychometric Best Practices

Practice 11: Conduct a psychometric study to demonstrate construct mapping, reliability and validity

Practice 12: Report the coefficient of reproducibility rather than alpha for a multi-item scale

The Facets of Measurement

The Facets of Measurement

Reliability and validity of a measure are constrained by the facets of measurement, namely the target population, the settings, and time

Classic example is IQ assessments and ethnicity

We need to ensure that evidence bearing on the psychometric properties of a measure match the facets of our research

If prior psychometric research does not map onto our facets, we may need to do a psychometric study

Psychometric Best Practices

Practice 11: Conduct a psychometric study to demonstrate construct mapping, reliability and validity

Practice 12: Report the coefficient of reproducibility rather than alpha for a multi-item scale

Practice 13: Ensure the psychometric evidence for a scale maps onto the facets of measurement in your research

Evaluating Extant Multi-Item Measures

Evaluating Extant Multi-item Measures

Is the scale relatively free of random error?

Is their validity data supporting the scale (is it correlated with constructs it should be correlated with and uncorrelated with constructs it should not be)?

Is the scale subject to social desirability/good impression bias

Have the above properties been established under conditions comparable to your research? Do the psychometric properties generalize?

Evaluating Extant Multi-item Measures

Are the individual items relatively "noise" free? Is there considerable unique variance in them? [examine item correlations – should be sizeable if considerable unique variance, consider modeling it]

Do the items meet the 10 criteria for (writing) good items?

Is the item metric precise enough and are the adverb qualifiers appropriate?

Are the orienting instructions appropriate (e.g., practice items, anchoring)?

Evaluating Extant Multi-item Measures

Is the scale true to its concept map? [Perform a confirmatory factor analysis – be wary of exploratory factor analysis]

(***Problems with EFA include (a) it is atheoretical, (b) issues of choosing number of factors, (c) issues of choosing a rotation method, (d) can not take into acccount small correlated errors, (e) unclear how to deal with violations of Thurstone simple structure coupled with factor score indeterminancy***)

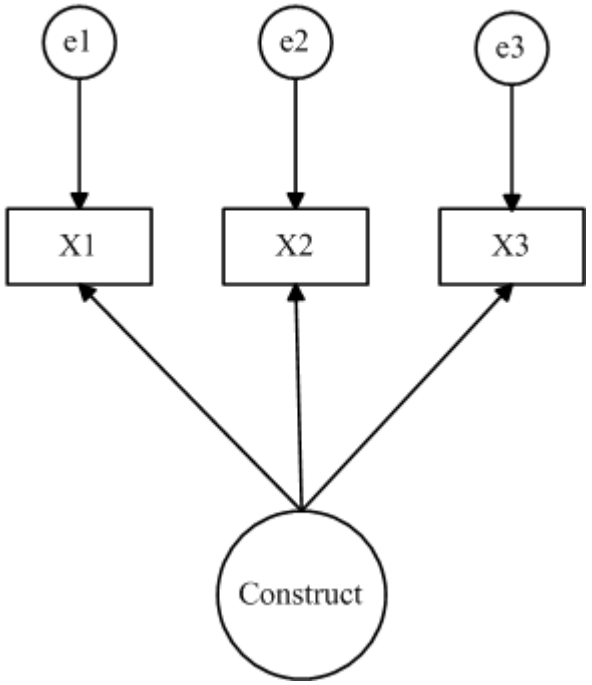
What is the composite reliability for the scale/sub-scales?

What data can I use to evaluate construct validity?

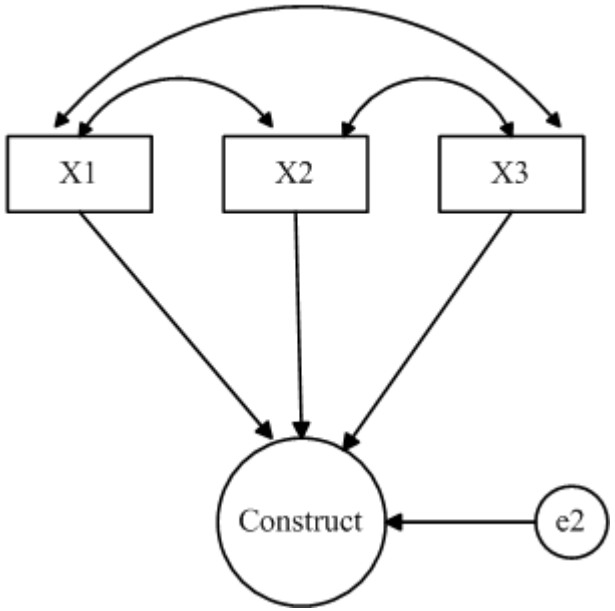
Reflexive versus Formative Models

Reflexive versus Formative Models

(a)



(b)



Reflexive = construct impacts "indicators"; Formative = construct is defined by the "indicators"[18-20]

Reflexive versus Formative Models

Examples include (a) social class (education, income, occupational prestige), (b) the human development index of countries (mortality rates, education, income), (c) exercise activity (running, sports, walking) and…. DSM 5 categories(?)

Does the presence of symptoms define the underlying construct or does the underlying construct impact symptoms?

Reflexive models assume high correlations among indicators and a unidimensional structure. Formative models do not

Reflexive versus Formative Models

DSM5 alcohol use dependency is characterized on a continuum from none to mild to moderate to severe (latent variable)

It is a combination of 11 diverse (dichotomous) symptoms (e.g., craving; failure to fulfill major obligations at home, work, or school; recurrent use in situations that are physically hazardous) in a formative sense

Some argue categories should be based on reflective models rather than formative models. Others argue the reverse

Additional Topics

Additional Topics

Arbitrary metrics

Measurement invariance

Ordinal level psychometric modeling

Item response theory

Psychometric Best Practices

Psychometric Best Practices

Practice 1: Adopt the six practices to minimize unreliability

Practice 2: Use SEM to gain perspectives on the impact of unreliability of measures on conclusions

Practice 3: Adopt the five practices to minimize social desirability/good impression influences on measures

Practice 4: Use SEM to gain perspectives on the impact of systematic measurement error on conclusions

Practice 5: Use strong rather than weak approaches to concept mapping

Psychometric Best Practices

Practice 6: Use the 10 strategies for writing items

Practice 7: Address issues of metric precision

Practice 8: Select adverb qualifiers and phrases that maximize interval level properties of the metric

Practice 9: Apply cognitive testing to refine items

Practice 10: Use anchoring

Psychometric Best Practices

Practice 11: Conduct a psychometric study to demonstrate construct mapping, reliability and validity

Practice 12: Report the coefficient of reproducibility rather than alpha for a multi-item scale

Practice 13: Ensure the psychometric evidence for a scale maps onto the facets of measurement in your research

Practice 14: Be clear about your underlying model (reflexive versus formative)

References

[1] Uziel, L. (2010). Rethinking social desirability scales : From impression management to interpersonally oriented self control. Perspectives on Psychological Science, 5, 243-252.

[2] Fleming, P. (2012). Social desirability, not what it seems: A review of the implications for self-reports. The International Journal of Educational and Psychological Assessment, 11, 3-22.

[3] Paulhus, D.L. (2002). Socially desirable responding: The evolution of a construct. In H. Braun, D.N. Jackson, & D.E. Wiley (Eds.), The role of constructs in psychological and educational measurement (pp. 67–88). Hillsdale, NJ: Erlbaum.

References

[4] Conway, J. & Lance, C. (2010). What reviewers should expect from authors regarding common method bias in organizational research. Journal of Business Psychology, 25, 325–334

[5] Podsakoff, P., MacKenzie, S. & Podsakoff, N. (2012). Sources of method bias in social science research and recommendations on how to control it. Annual Review of Psychology, 63, 539-569

[6] Rorer, L. (1965). The great response-style myth. Psychological Bulletin, 63, 129-156.

References

[7] Wiggins, J.S. (1973). Personality and prediction: Principles of personality assessment. Reading, MA: Addison-Wesley

[8] Wang, L., Miller, M., Schmitt, M. & Wen, F. (2013). Assessing readability formula differences with written health information materials: Application, results, recommendations. Research in Social and Administrative Pharmacy, 9, 503-516

[9] Burke, V. & Greenberg, D. (2010). Determining readability: How to select and apply easy-to-use readability formulas to assess the difficulty of adult literacy materials. Adult Basic Education and Literacy Journal, 4, 34-42.

References

[10] Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized data. American Sociological Review, 46, 232-239.

[11] Lietz, P. (2010). Research into questionnaire design: A summary of the literature. International Journal of Market Research, 52, 249-272

[12] Beckstead, J. (2014). On measurements and their quality. Paper 4: Verbal anchors and the number of response options in rating scales. International Journal of Nursing Studies, 51, 807–814.

References

[13] Woltz, D., Gardner, M., Kircher, J. & Burrow-Sanchez, J. (2012). Relationship between perceived and actual frequency represented by common rating scale labels, Psychological Assessment. 24, 995–1007

[14] Spector, P. (1976). Choosing response categories for summated rating scales, Journal of Applied Psychology, 61, 374-375.

[15] Kingstrom & Bass. (1981). A critical analysis of studies comparing behaviorally anchored ratings scales (BARS) and other rating formats. Personnel Psychology, 34, 263–89.

References

[16] Wedell, D., Parducci, A. & Lane, M. (1990). Reducing the dependence of clinical judgment on the immediate context: effects of number of categories and type of anchors. Journal of Personality ands Social Psychology. 58, 319-329.

[17] Graham, J. (2005). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them Educational and Psychological Measurement, 66, 930-944.

[18] Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. Psychological Bulletin, 100, 305-314

References

[19] MacCoun (2013). The puzzling unidimensionality of DSM-5 substance use disorder diagnoses. Frontiers in Psychiatry, 4, 1-5.

[20] Martin, C. (2013). The puzzling unidimensionality of the DSM substance use disorders: Commentary. Frontiers in Psychiatry, 4, 1