Methodological Issues in Measuring the Development of Character

Noel A. Card
Department of Human Development and Family Studies
College of Liberal Arts and Sciences

Supported by a grant from the John Templeton Foundation (IF#47910)

UCONN

Preliminary – my background

- Ph.D. in Clinical Psychology
- Research in Developmental Psychology / Science
  - Primarily study child & adolescent aggression & peer relationships
  - Relatively new to study of character development
- Postdoc in Quantitative Psychology
- Academic positions in
  - Human Development and Family Studies
  - Measurement, evaluations, and assessment

Preliminary – my background

- Ph.D. in Clinical Psychology
- Research in Developmental Psychology / Science
  - Primarily study child & adolescent aggression & peer relationships
  - Relatively new to study of character development
- Postdoc in Quantitative Psychology
- Academic positions in
  - Human Development and Family Studies
  - Measurement, evaluations, and assessment

Outline for talk

- Fundamental psychometric properties of good measures
- Relevance to the study of character development
- Two examples of different situations
- Suggestions

Fundamental properties

- Reliability
- Validity
- Equivalence

Fundamental properties

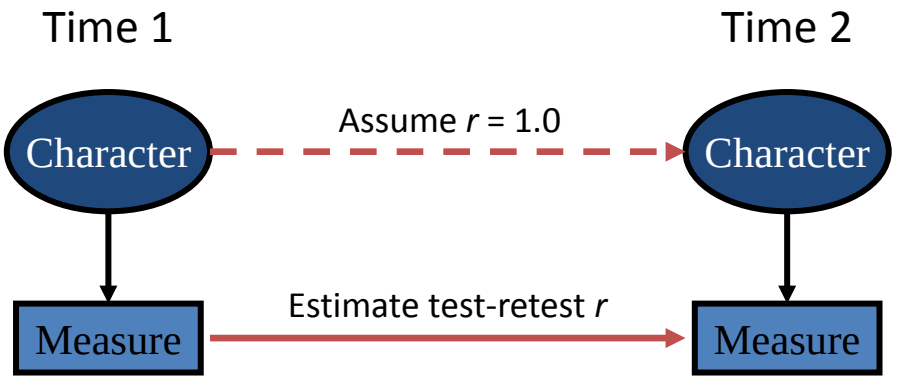
- Reliability = repeatability of multiple measures of a construct

Fundamental properties

- Reliability = repeatability of multiple measures of a construct
  - Internal consistency reliability
    - Repeatability across multiple items of a scale
    - Typically assessed with Cronbach's α
      - Ranges from 0 to 1, with higher values → higher reliability
      - Assumes 'parallel items' (i.e., all items have same variances and same correlations with total score) but this assumption is rarely tested
    - Other indices do not make these assumptions (e.g., McDonald's Ω)

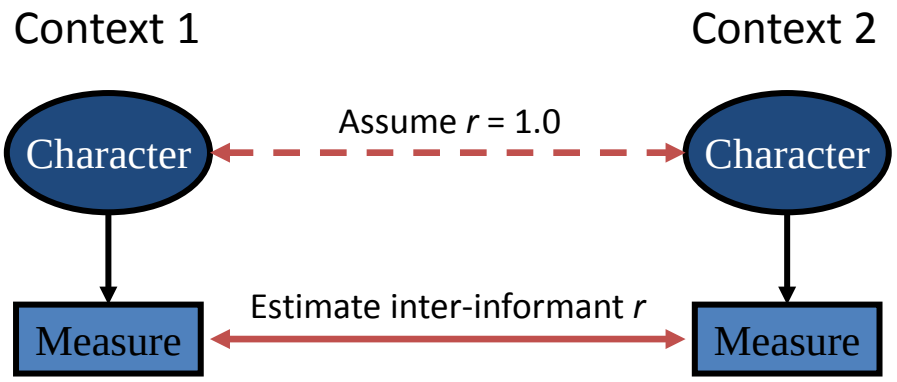
Fundamental properties

- Reliability = repeatability of multiple measures of a construct
  - Internal consistency reliability
  - Test-retest reliability
    - Repeatability across multiple measurement occasions
    - Assumes that construct is stable over time span
      - Time span must be short enough to avoid developmental / intervention instability



Fundamental properties

- Reliability = repeatability of multiple measures of a construct
  - Internal consistency reliability
  - Test-retest reliability
  - Inter-informant reliability
    - Repeatability across multiple reporters
    - Assumes that construct is stable over contexts of observation



Fundamental properties

- Summary regarding reliability
  - Three forms:
    - Internal consistency reliability
    - Test-retest reliability
    - Inter-informant reliability

Fundamental properties

- Summary regarding reliability
  - Three forms
  - In practice:
    - Internal consistency (with α) commonly considered b/c:
      - Necessary to have adequate α or use latent variable analyses
      - Multi-item scales commonly used; repeated-measures or multiple informants are not

Fundamental properties

- Summary regarding reliability
  - Three forms
  - In practice
  - Some cautions:
    - All reliabilities are *estimates* from a sample (not a property of the measurement tool).
    - Might vary by age, context, intervention condition, etc.
    - Do not overemphasize reliability:
      - Meaningful instability across occasions, development, or contexts
      - Need to give as much (more?) attention to validity and equivalence

Fundamental properties

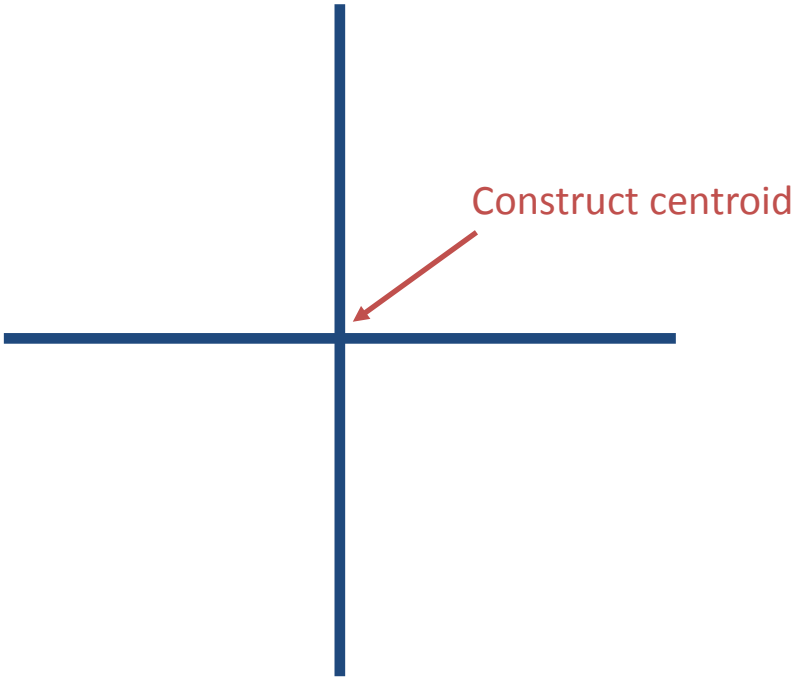
- Reliability = Repeatability of multiple measures of a construct
- Validity = Extent that the measurement instrument assesses what we intend it to measure

Fundamental properties

- Validity = Extent that the measurement instrument assesses what we intent it to measure
  - Example: Does a measure of prosocial behavior capture individual differences in frequency of prosocial behavior?
  - Versus:
    - A specific subdomain of the construct (e.g., helping teacher)
    - An irrelevant construct (e.g., social desirability, academic achievement)
  - Requires a clear definition of the construct
    - Recall Larry Nucci's talk, reflections, and discussion yesterday

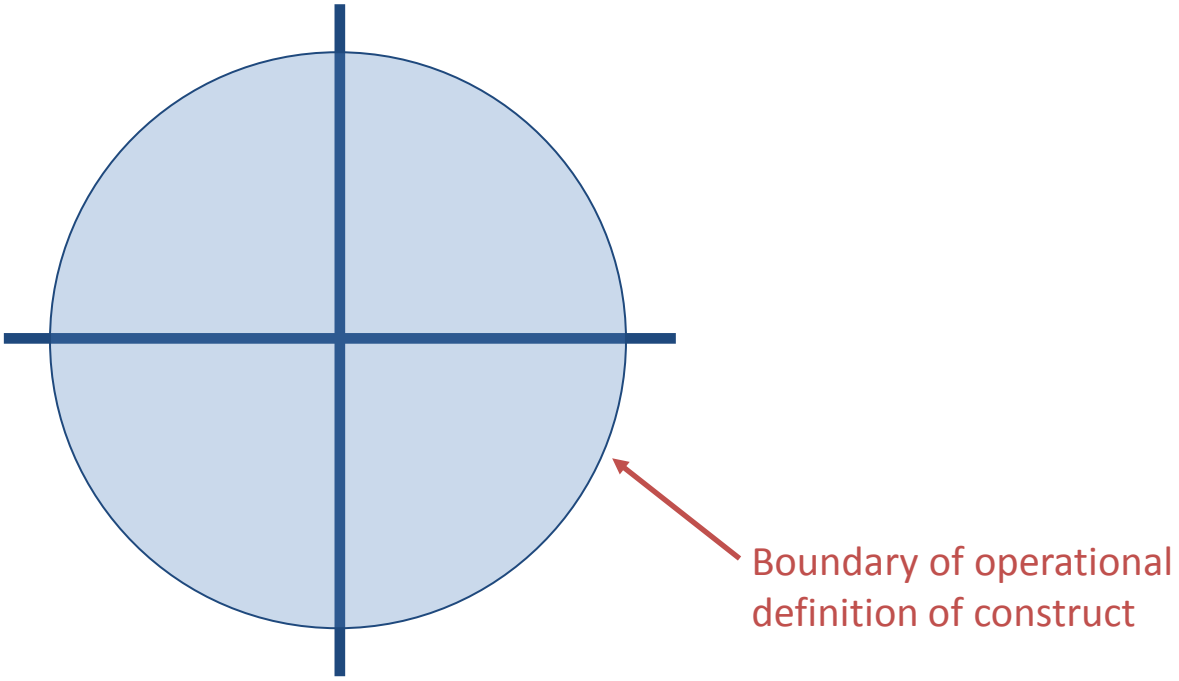
Fundamental properties

- Validity = Extent that the measurement instrument assesses what we intent it to measure
  - Domain representative framework (Nunnally, 1978):
    - Note: We can use this framework without assuming immutability of character or reducing human complexity to a single variable



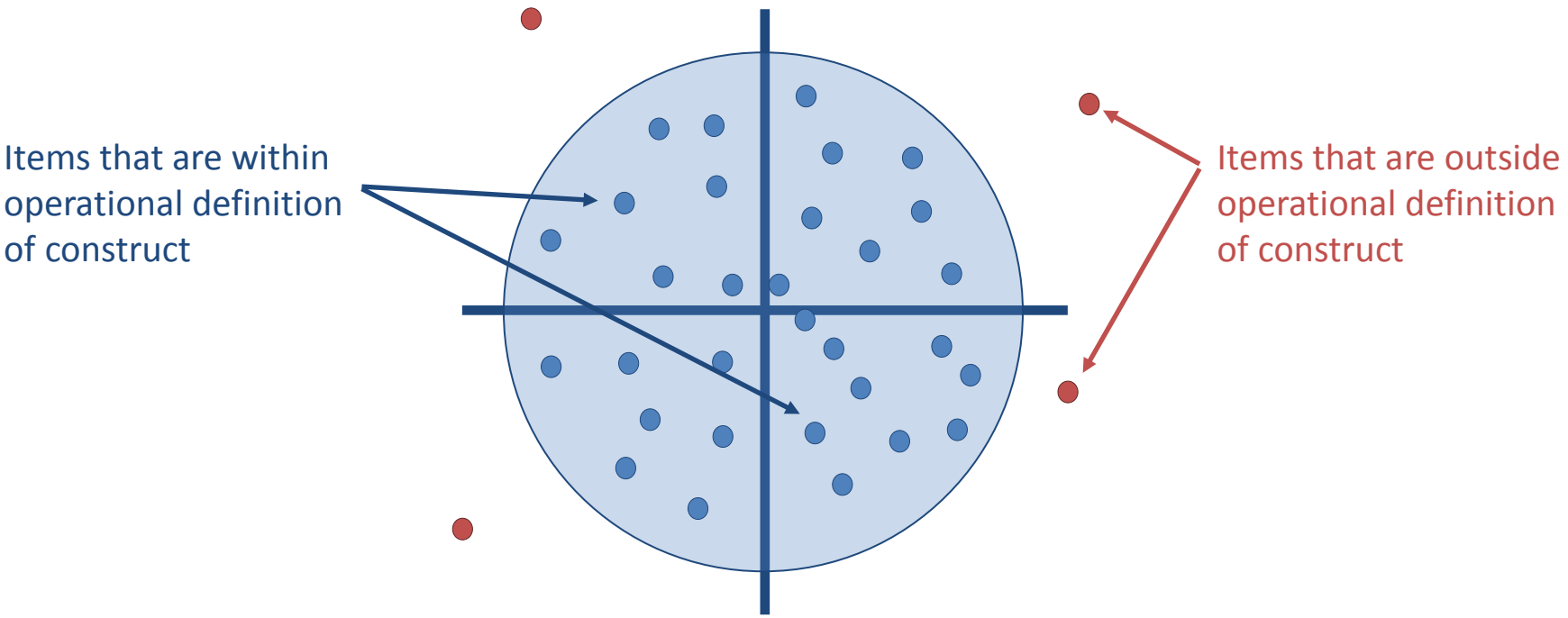
Fundamental properties

- Validity = Extent that the measurement instrument assesses what we intent it to measure
  - Domain representative framework (Nunnally, 1978):



Fundamental properties

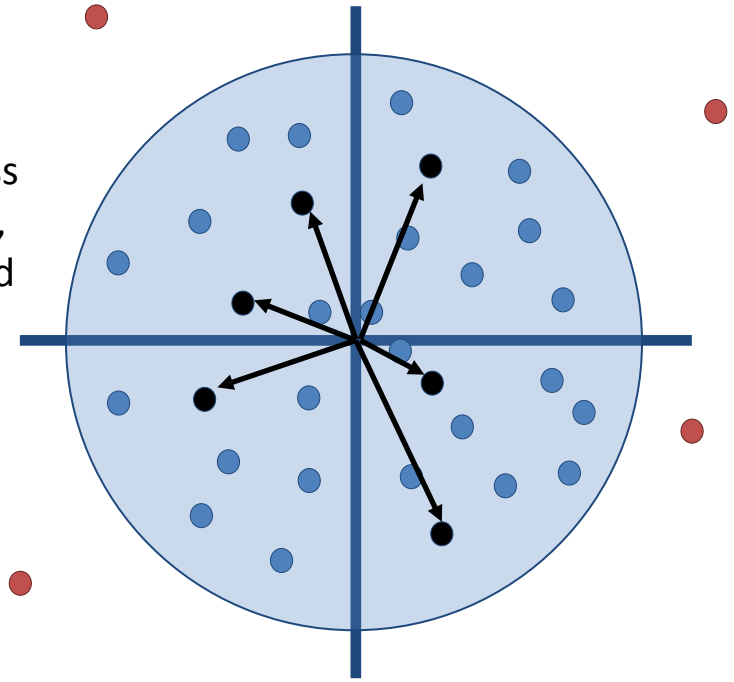
- Validity = Extent that the measurement instrument assesses what we intent it to measure
  - Domain representative framework (Nunnally, 1978):



Fundamental properties

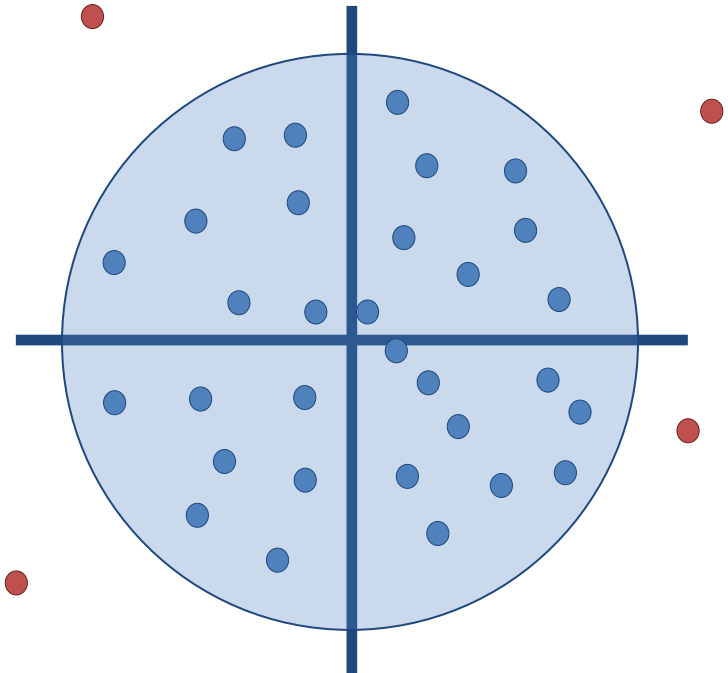
- Validity = Extent that the measurement instrument assesses what we intent it to measure
  - Domain representative framework (Nunnally, 1978):

Composite scores of items across construct space will, on average, triangulate on construct centroid



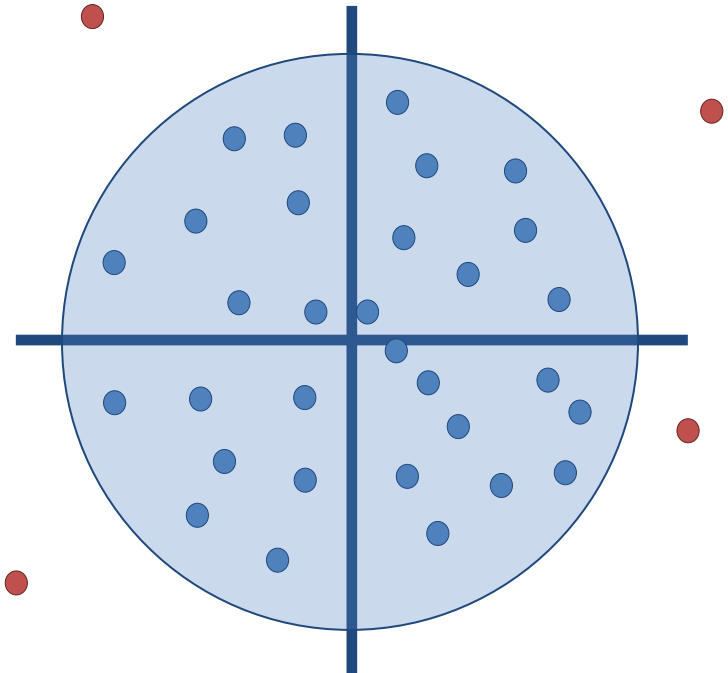
Fundamental properties

- Validity
  - The danger of prioritizing reliability over validity:



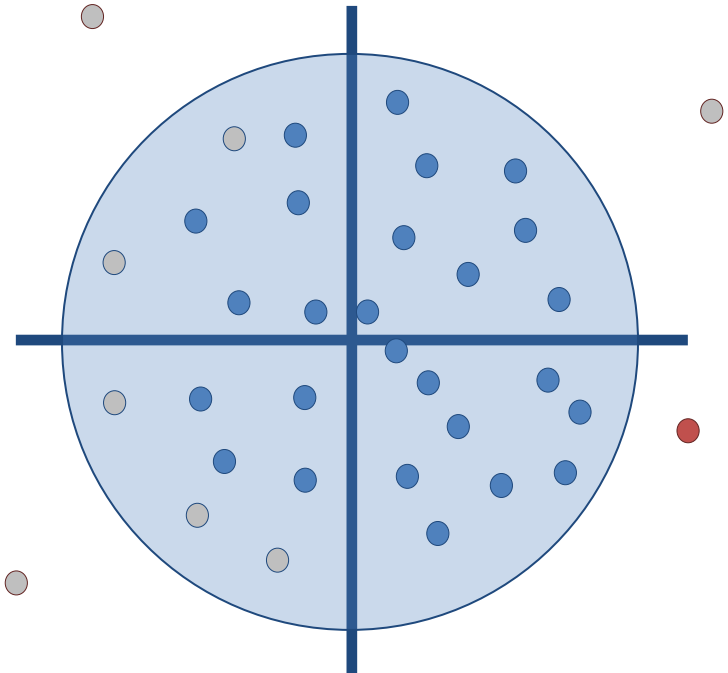
Fundamental properties

- Validity
  - The danger of prioritizing reliability over validity:
    - The initial set of items across the domain is diverse (low inter-item *r*s) and might have low reliability



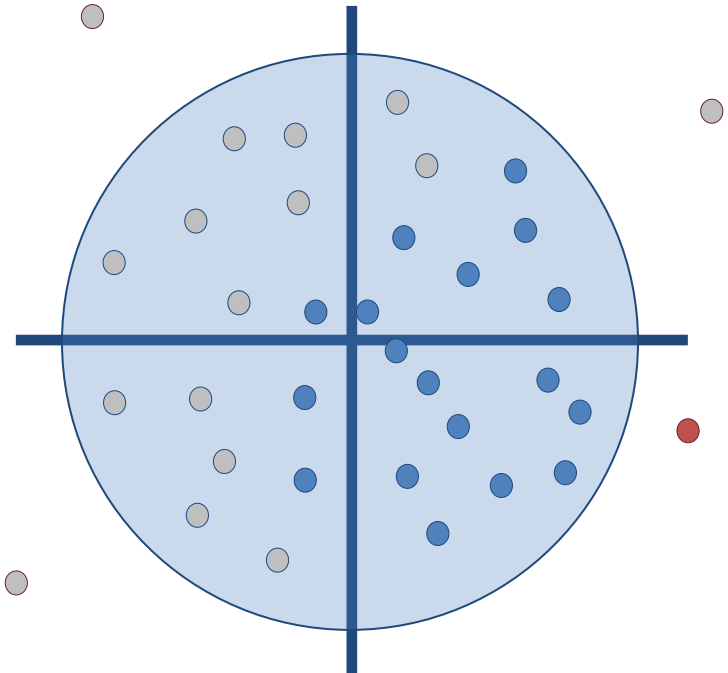
Fundamental properties

- Validity
  - The danger of prioritizing reliability over validity:
    - The initial set of items is diverse and might have low reliability
    - Efforts to remove items to improve reliability…



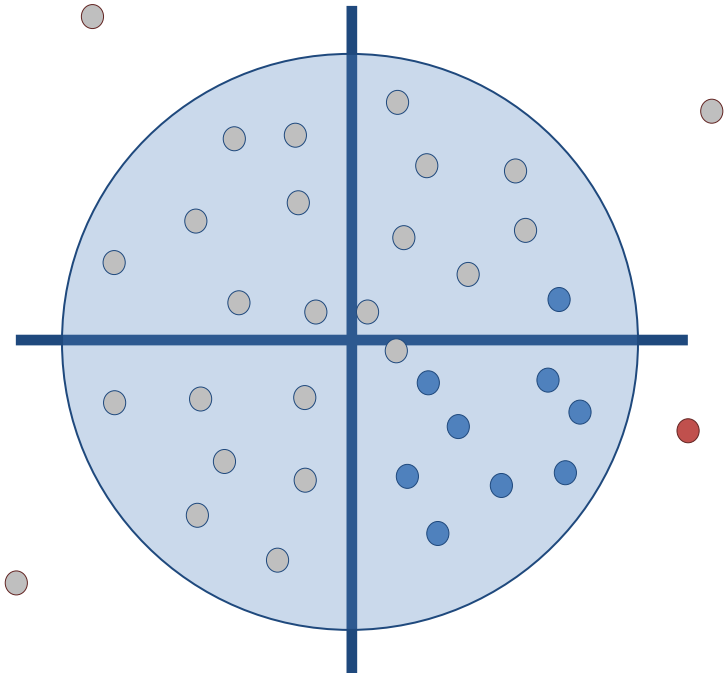
Fundamental properties

- Validity
  - The danger of prioritizing reliability over validity:
    - The initial set of items is diverse and might have low reliability
    - Efforts to remove items to improve reliability…



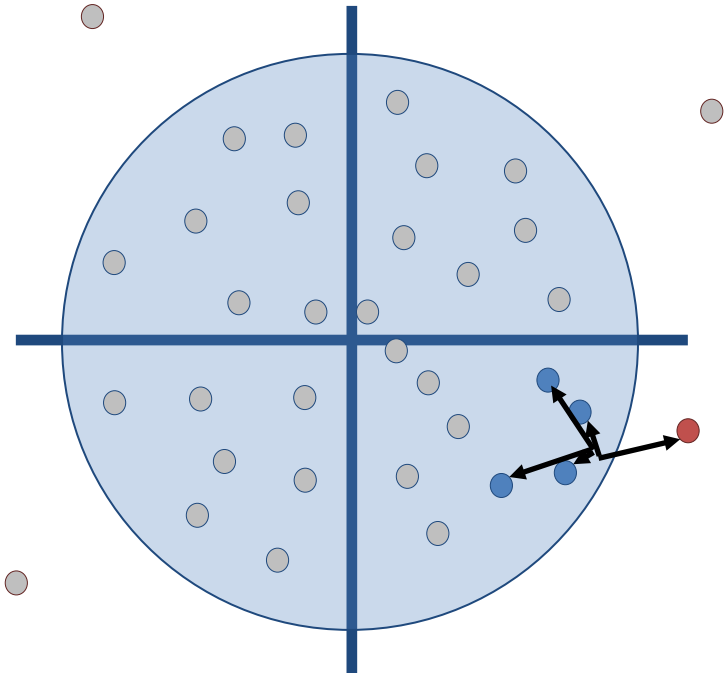
Fundamental properties

- Validity
  - The danger of prioritizing reliability over validity:
    - The initial set of items is diverse and might have low reliability
    - Efforts to remove items to improve reliability…



Fundamental properties

- Validity
  - The danger of prioritizing reliability over validity:
    - The initial set of items is diverse and might have low reliability
    - Efforts to remove items to improve reliability…
    - …leads to a reliable measure <u>of the wrong thing</u>



Fundamental properties

- Summary regarding validity
  - Validity = Extent that the measurement instrument assesses what we intent it to measure
  - Requires clear operational definition of character
    - Validity may depend on theory / perspective
    - Needs to sensitive to change over development
    - Needs to be sensitive to change across intervention

Fundamental properties

- Reliability = Repeatability of multiple measures of a construct
- Validity = Extent that the measurement instrument assesses what we intent it to measure
- Equivalence = A measurement instrument performs in the same way across situations…
  - AKA measurement equivalence, measurement invariance, factorial equivalence, factorial invariance, (absence of) differential item functioning

Fundamental properties

- Equivalence = A measurement instrument performs in the same way across situations…
  - Groups:
    - Males versus females
    - Ethnic groups
    - Treatment versus control
  - Time:
    - Pre- versus post-intervention
    - Multiple waves of a longitudinal study

Fundamental properties

- Three levels of equivalence
  - Configural = Same items load onto same constructs (technically: Same pattern of fixed and free parameters)

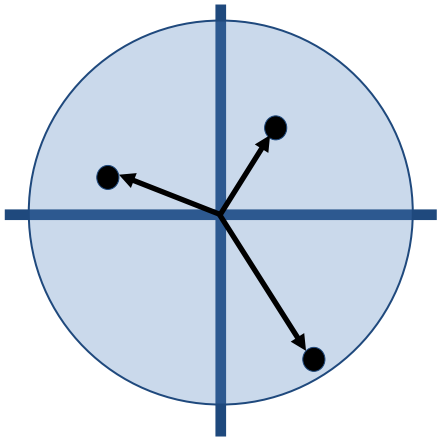
Fundamental properties

- Three levels of equivalence
  - Configural = Same items load onto same constructs
  - Weak (AKA metric, loading): Same relative strengths of factor loadings

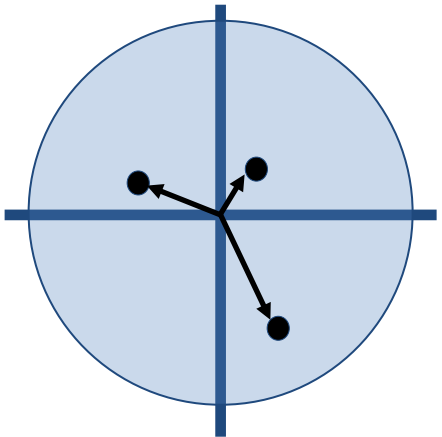
Fundamental properties

- Three levels of equivalence
  - Configural = Same items load onto same constructs
  - Weak (AKA metric, loading): Same relative strengths of factor loadings
    - Ensures construct centroid defined equivalently
    - Allows comparisons of variances and correlations across groups / time
    - Allows meaningful estimate of (inter-individual) stability

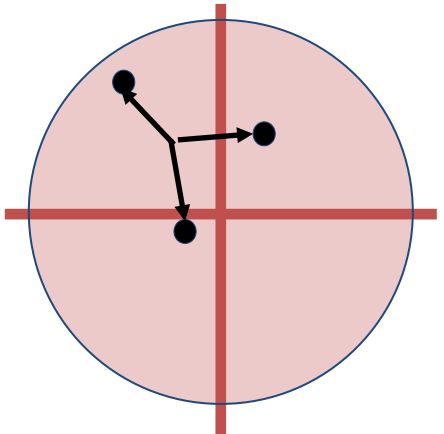
Group / Time 1



Group / Time 2



Noninvariance

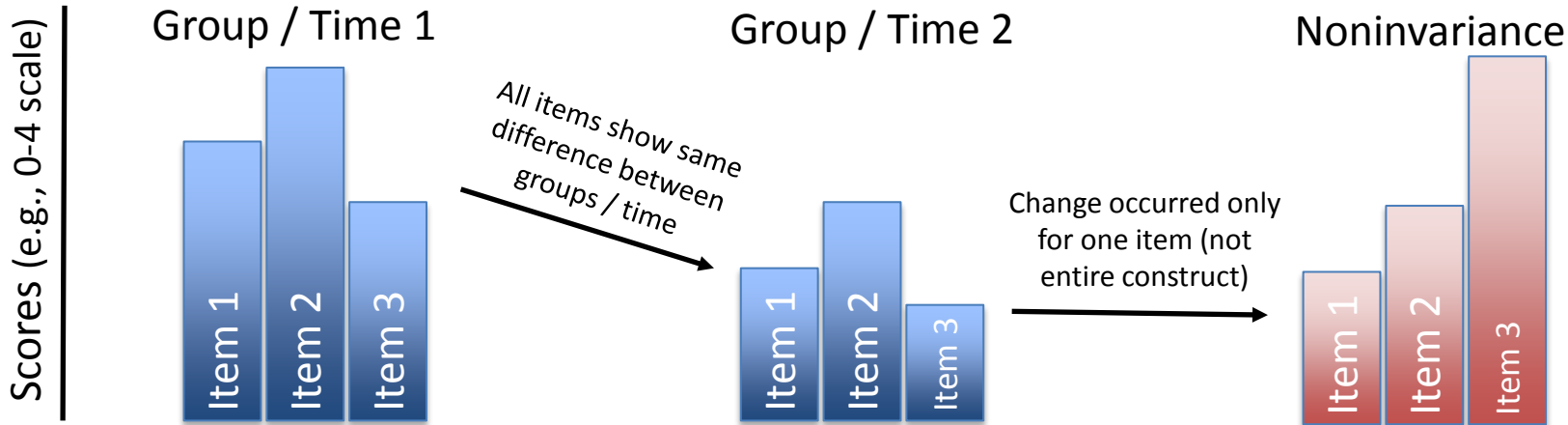


Fundamental properties

- Three levels of equivalence
  - Configural = Same items load onto same constructs
  - Weak: Same relative strengths of factor loadings
  - Strong (AKA scalar, intercept): Same relative magnitudes of item means (technically indicator intercepts)

Fundamental properties

- Three levels of equivalence
  - Configural = Same items load onto same constructs
  - Weak: Same relative strengths of factor loadings
  - Strong (AKA scalar, intercept): Same relative magnitudes of item means (technically indicator intercepts)
    - Ensures construct means defined equivalently
    - Allows comparisons of means across groups / time
    - Allows meaningful estimate of (intra-individual) stability



Fundamental properties

- Summary regarding equivalence
  - Equivalence = A measurement instrument performs in the same way across groups and/or time
  - Three levels:
    - Configural
    - Weak
    - Strong
  - Frequency of testing:
    - Very rarely (in areas of character / character development I have read)
  - Challenges:
    - Requires use on Confirmatory Factor Analysis (CFA; or related techniques like IRT)

Roadmap

- Fundamental psychometric properties of good measures
- Relevance to the study of character development
- Two examples of different situations
- Suggestions

Roadmap

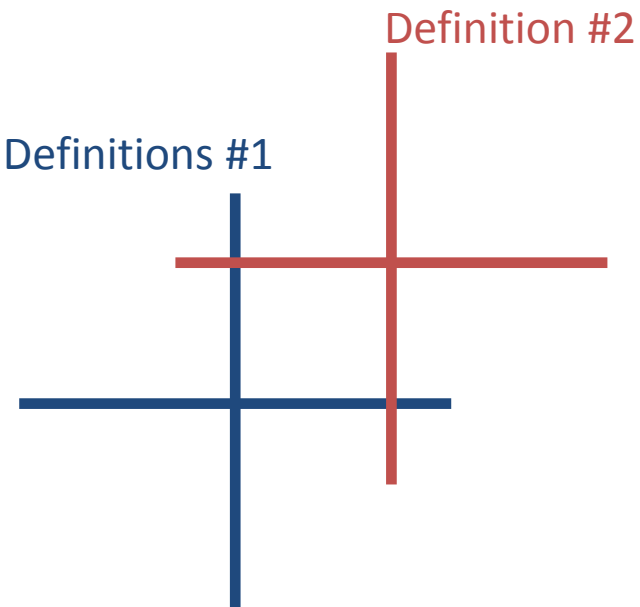
- Fundamental psychometric properties of good measures
- Relevance to the study of character development
- Two examples of different situations
- Suggestions

Relevance

- Relevance of high quality measurement to the study of character & character development, related to…
  - Definitions
  - Populations and contexts
  - Study designs

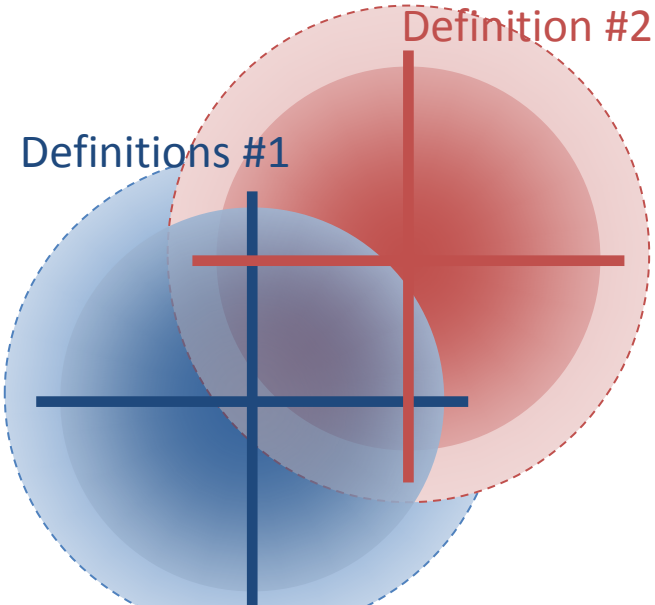
Relevance

- Definitions
  - Multiple definitions of constructs



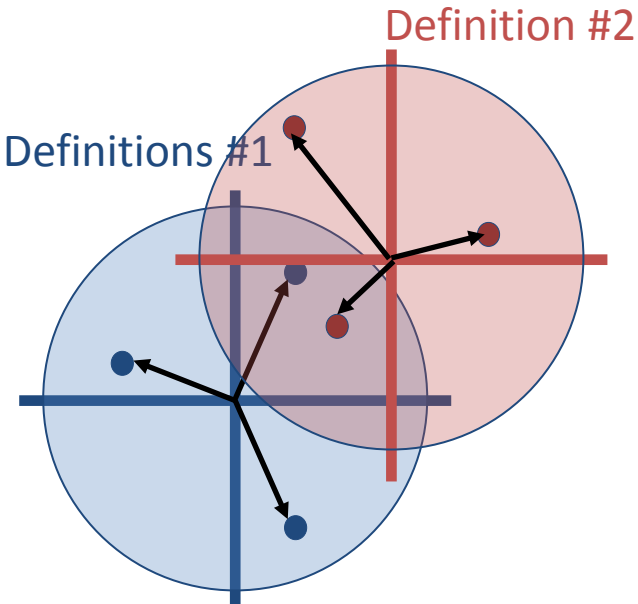
Relevance

- Definitions
  - Multiple definitions of constructs
  - Fuzzy boundaries of operational definitions



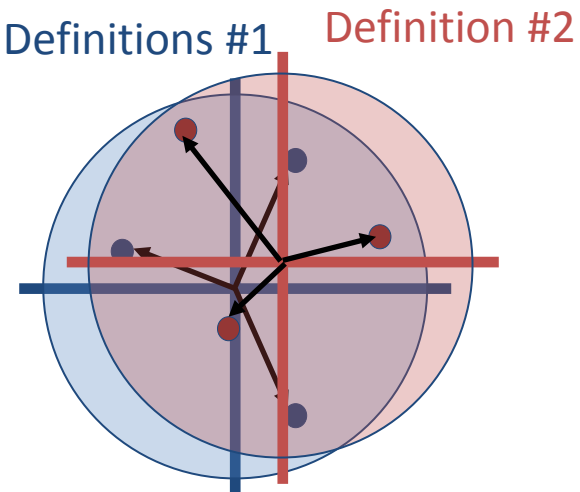
Relevance

- Definitions
  - Multiple definitions of constructs
  - Fuzzy boundaries of operational definitions
  - Is it possible to identify common measurement tools?
    - Can we at least identify common items (for e.g., large scale datasets, secondary data analysis)?



Relevance

- Definitions
  - Multiple definitions of constructs
  - Fuzzy boundaries of operational definitions
  - Is it possible to identify common measurement tools?
  - Should we aim to align definitions?
    - Is theoretical diversity or methodological similarity more important?



Relevance

- Populations and contexts
  - Field is marked (or should be) by attention to diversity in sampling…
    - Populations studied (e.g., gender, ethnicity)
    - Contexts (e.g., school, after-school, 4-H, scouting, home)
    - Language and culture (e.g., cross-national)
    - Age (character *development* is inherently interested in change across time)

Relevance

- Populations and contexts
  - Field is marked by attention to diversity in sampling
  - Field is marked by limited attention to estimating psychometric properties across these diverse populations and contexts
    - Must assess psychometric properties in every study
    - Explicit attention to evaluating measurement equivalence

Relevance

- Study designs
  - Basic designs:
    - Naturalistic (concurrent and longitudinal) studies
    - Experimental (or quasi-experimental) intervention studies
      - Recall talks by Berkowitz, Durlak, and Trochim, reflections, and discussion yesterday

Relevance

- Study designs
  - Intervention studies must also (especially?) consider equivalence
    - Interventions, programs, and policies may change the measurement of character across time and/or group
  - Failure to establish equivalence leads to any of these scenarios:
    - Intervention increases (e.g.,) prosocial behavior and does not impact measurement
    - Intervention leads to higher measured prosocial behavior but no real changes in the construct (e.g., socially desirable responding)
    - Intervention leads to higher prosocial and impacts the measurement so intervention effect is exaggerated
    - Intervention leads to higher prosocial and impacts the measurement so intervention effect is hidden
    - Intervention *reduces* prosocial behavior but heightens reporting, obscuring the harmful impact of the intervention

Relevance

- Study designs
  - Intervention studies must also (especially?) consider equivalence
    - Interventions may change the measurement of character across time and/or group
  - Failure to establish equivalence leads to any of these scenarios:
  - Bottom line: We cannot have confidence in intervention effects without ensuring that they impact *character* rather than *measurement* of character

Roadmap

- Fundamental psychometric properties of good measures
- Relevance to the study of character development
- Two examples of different situations
- Suggestions

Roadmap

- Fundamental psychometric properties of good measures
- Relevance to the study of character development
- Two examples of different situations
- Suggestions

Two examples

- Two examples
  - From ongoing meta-analysis synthesizing psychometric properties of 11 character strengths
    - Funded by John Templeton Foundation (ID#47910)
  - Two examples (intended to represent many areas of character development research):
    - Gratitude: Limited number of widely-used measurement instruments
    - Humility: Absence of widely-used measurement instruments

Two examples

- Gratitude
  - Operational definition:
    - Sense of thankfulness or appreciation in response to receiving a gift, whether that gift is a tangible object given by someone else, experiences that one has had in life, or positive characteristics such as one's health (e.g., Peterson & Seligman, 2004)

Two examples

- Gratitude
  - Operational definition:
    - Sense of thankfulness or appreciation in response to receiving a gift, whether that gift is a tangible object given by someone else, experiences that one has had in life, or positive characteristics such as one's health (e.g., Peterson & Seligman, 2004)
  - Small number of widely-used instruments. Two of these:
    - GQ-6 (McCullough, Emmons, & Tsang, 2002): Six-item gratitude questionnaire
    - GRAT (Watkins, Woodward, Stone, & Kolts, 2003): Gratitude Resentment and Appreciation Test

Two examples

- Gratitude
  - GQ-6 (McCullough et al., 2002):
    - Study 1: First psychometric analyses
      - 39 item measure administered to college undergrads
      - EFA indicated one factor
      - Authors trimmed to 6 items based on both item-total *r* and conceptual criteria
      - Correlations with other reporters (inter-informant reliability) and other self-report measures expected to correlate (construct validity)
    - Study 2: Broader sample of adults
      - Administered 6 item (trimmed) questionnaire to wider age span of adults
      - Similar evidence of construct validity
    - Studies 3 and 4 addressed substantive questions
      - Used college undergrads
      - No specific focus on psychometric properties

Two examples

- Gratitude
  - GRAT (Gratitude Resentment and Appreciation Test; Watkins et al., 2003):
    - Study 1: First psychometric analyses
      - 55 item measure administered to college undergrads
      - 9 items dropped to improve internal consistency
      - Expected four factor solution, but EFA indicated three factors
    - Study 2:
      - A second sample of college undergrads
      - Assumed three factor solution found in study 1
      - Evaluated test-retest reliability and evidence of construct validity
    - Studies 3 and 4
      - Experimental manipulation to impact gratitude (3 factors)

Two examples

- Gratitude
  - Critiques of these two seminal studies
    - Strengths:
      - Impressive translations of theoretically-grounded conceptualization of gratitude into tractable measures
      - Collectively, the 8 studies examined many of the psychometric properties:
        - » Factor structure
        - » Internal consistency reliability
        - » Inter-informant reliability
        - » Test-retest reliability
        - » Many correlations informing validity

Two examples

- Gratitude
  - Critiques of these two seminal studies
    - Strengths:
    - Limitations:
      - Both initial studies removed items to improve reliability (though McCullough et al also gave conceptual consideration)
      - Both studies used decision about items to retain (and factor structure, in Watkins et al., 2003) in subsequent studies without replication
      - 7 of 8 samples were undergrads
        - » The one non-college sample was not ethnically diverse (91% White)
      - Validity evidence drawn primarily from self-report measures without considering shared-method variance
      - Neither paper reported results of measurement equivalence (across e.g., gender, experimental manipulations of Studies 3 & 4 of Watkins et al.)

Two examples

- Gratitude
  - Critiques of these two seminal studies
    - Strengths:
    - Limitations:
    - Conclusions:
      - These seminal papers do not need to be definitive
      - We should be aware of limitations that need to be addressed in subsequent studies
      - Should not view measures as definitively supported.
        - » Needs to be ongoing evaluation of psychometric properties
        - » Possibility for modifying instruments for particular populations, contexts, or applications

Two examples

- Gratitude
  - Small number of widely-used instruments
    - Literature review identified 108 studies using at least one of four measures
  - Many advantages of this situation:
    - Even if many studies are individually homogeneous, the collection of studies is diverse (in population, context, and application)
    - Wealth of previous studies the researchers can refer to identify an acceptable measure for a particular use
      - However, reporting of full range of psychometric properties is frequently lacking

Two examples

- Humility
  - Operational definition:
    - Character strength that includes having an accurate sense of one's abilities and achievements, an ability to acknowledge mistakes, openness to advice and new ideas (e.g., Peterson & Seligman, 2004)

Two examples

- Humility
  - Operational definition:
    - Character strength that includes having an accurate sense of one's abilities and achievements, an ability to acknowledge mistakes, openness to advice and new ideas (e.g., Peterson & Seligman, 2004)
  - Absence of widely-used measures
    - Seemingly every researcher develops a unique measure for each study

Two examples

- Humility
  - Absence of widely-used measures causes challenges:
    - Ambiguity if construct is studied
      - Reader must have operational definition and decide if study measured humility
      - Some authors used “humility” for different constructs (outside operational definition)
      - Some authors used different terms for this construct

Two examples

- Humility
  - Absence of widely-used measures causes challenges:
    - Ambiguity if construct is studied
    - Difficult to use prior literature to guide selection of measures
      - Insufficient use of any single measure to know the situations (populations, contexts, etc) in which it performs well
      - A researcher might find study of closest population, context, etc., but the measure might not match the researcher's operational (theoretical) definition
      - Or, a researcher might use a measure matching the desired operational definition, but only hope that it works well in the current situation
        - » Practice supported by false beliefs that psychometric properties are properties of the measurement instrument
      - Or, just do not rely on prior literature to guide measurement selection

Two examples

- Humility
  - Absence of widely-used measures causes challenges:
    - Ambiguity if construct is studied
    - Difficult to use prior literature to guide selection of measures
    - Virtually impossible to synthesize any study results
      - Cannot have efficient accumulation of empirical knowledge about a construct
        - » What are best ways of measuring?
        - » What are correlates?
        - » What are most effective programs?

Two examples

- Summary
  - Two examples (intended to represent many areas of character development research):
    - Gratitude: Limited number of widely-used measurement instruments
    - Humility: Absence of widely-used measurement instruments
  - The former situation is better than the latter
    - But, neither is ideal (or, without cautions)

Roadmap

- Fundamental psychometric properties of good measures
- Relevance to the study of character development
- Two examples of different situations
- Suggestions

Roadmap

- Fundamental psychometric properties of good measures
- Relevance to the study of character development
- Two examples of different situations
- Suggestions

Suggestions

- Suggestions for:
  - Study planning
  - Results reporting
  - Synthesis of research

Suggestions

- Study planning
  - Researchers as thoughtful consumers of prior research
    - Actively read and evaluate large amount of prior research with goal of selecting good measurement instruments
    - Consideration of multiple aspects of psychometric quality
    - Recognition that psychometric properties are population, context, & research demand specific
      - E.g., a good measure of character might not be a good measure of character *development*

Suggestions

- Study planning
  - Researchers as thoughtful consumers of prior research
  - Researchers are empowered to modify existing measurement instruments
    - Rigid application of existing scales neglects knowledge of research setting
    - Might be paired with ongoing qualitative studies or mixed-methods scale development research

Suggestions

- Study planning
  - Researchers as thoughtful consumers of prior research
  - Researchers are empowered to modify existing measures
  - Value in multidisciplinary teams in selecting measures

Suggestions

- Reporting findings
  - Full reporting of psychometric properties
    - Internal consistency: Cronbach's α plus other indices
    - Other forms of reliability (test-retest, inter-informant agreement) if available
    - Validity evidence (sometimes blurred with substantive questions)
    - Equivalence testing (for any groups, settings, or measurement occasions that could plausibly change measurement)

Suggestions

- Reporting findings
  - Full reporting of psychometric properties
  - Challenges
    - Some analyses (e.g., measurement equivalence) are technically demanding
      - Is time / effort / consultation available?
    - Are journals / publication outlets willing to dedicate space to psychometric results?

Suggestions

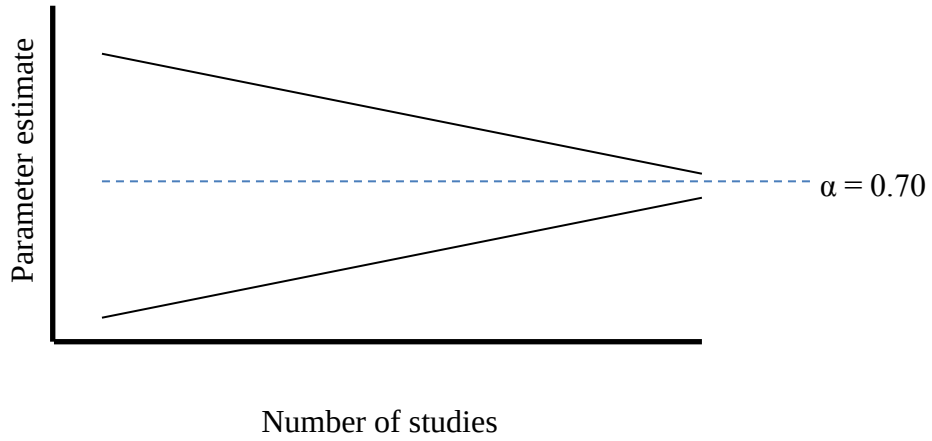
- Reporting findings
  - Full reporting of psychometric properties
  - Challenges
  - Studies focused on psychometrics
    - Some studies (or some aspects of studies) devoted to psychometric results
      - Evaluating of full psychometric results
      - Diverse populations, contexts, methods of measurement
      - Direct comparison of multiple measurement instruments
    - Could be built into planned analyses of a larger study
    - Need to shift perceptions in field so psychometric results are highly valued (versus just a preliminary to more interesting results)

Suggestions

- Research accumulation
  - Meta-analysis as a tool
    - Meta-analysis is a methodology for systematically search, coding, and analyzing existing research results (including psychometric properties)
    - Some advantages over primary study focus on psychometrics:
      - Larger overall sample
      - Greater diversity (e.g., countries, research settings)
      - Incorporating many measures

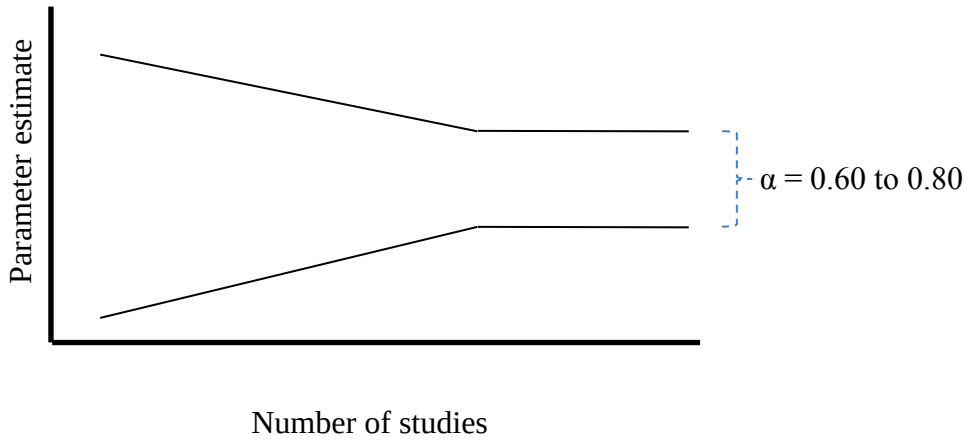
Suggestions

- Synthesis of exact replications
  - Each additional study provides more precise estimate of psychometric properties
  - Applies when all studies use same sample, measure, methodology, etc.



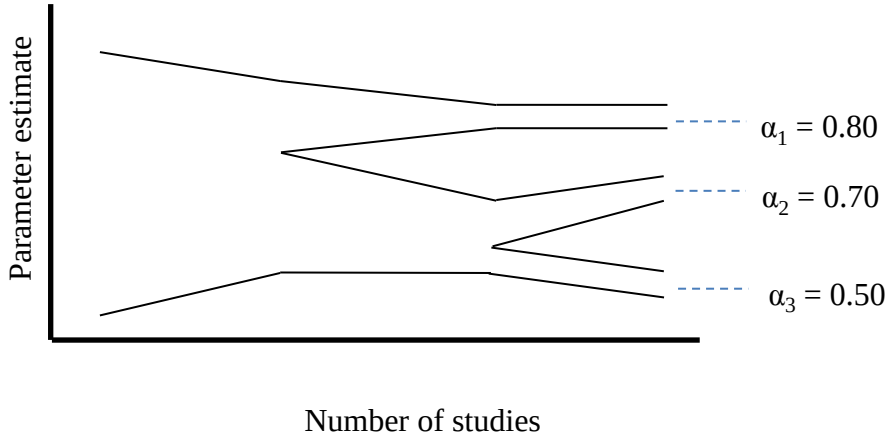
Suggestions

- Synthesis of unplanned inexact replications
  - Each additional study provides more precise estimate of psychometric properties up to a point, then precision is limited by unknown / unanalyzed differences across studies
  - Likely most common, because studies differ in many ways (e.g., many differences in samples, ages, context, measurement approaches, research demands)



Suggestions

- Synthesis of planned inexact replications
  - When further studies do not provide further increases in precision, then…
    - … systematically code plausible differences in existing studies, or
    - … conduct planned (intentional) inexact replications varying samples, contexts, measures, methodologies, etc.
  - Allows precision to inform specific types of future studies



Suggestions

- Synthesis of existing studies
  - Advantages:
    - Provide information based on all existing studies
      - High power / precision
      - Diversity of samples, methodologies, etc
    - Rely on well-established methodological / statistical practices
      - Effectively processes a lot of information
      - Avoids subjectivity

Suggestions

- Synthesis of existing studies
  - Advantages
  - Challenges:
    - Requires adequate number of studies using same / similar measures
    - Requires consistent reporting of psychometric properties
    - Existing studies should be of sufficient quality to meaningfully combine
    - Existing studies should have variability (e.g., in sample ethnicities, age) to identify variability in psychometric properties
      - Vast majority in my ongoing review of character strengths studied adult samples
      - Vast majority were concurrent (unclear if sensitive to change)

Suggestions

- Summary of suggestions
  - Change is needed in all three areas:
    - Study planning
    - Results reporting
    - Synthesis of research

Coda

- Covered four topics:
  - Fundamental psychometric properties of good measures
  - Relevance to the study of character development
  - Two examples of different situations
  - Suggestions

Coda

- Covered four topics
- Acknowledge no easy solution
  - Needs a shift in attention and valuation of good measurement

Coda

- Covered four topics
- Acknowledge no easy solution
- Limits to my perspective
  - Focused on quantitative measurement of individual differences
    - Versus cohesion / person-centered
  - Neglected other quantitative topics
    - Growth / change across development
    - Causal / predictive relations of character with other outcomes
    - How to conceptualize individual differences in character (discrete vs. continuous)
  - Neglected qualitative and mixed-methods approaches

- Please contact for questions / comments:
  - noel.card@uconn.edu

- Thanks to John Templeton Foundation for supporting my work in this area: Card, N. A. (in progress). What is known about existing measures: Meta-analyses of psychometric properties of measures of character strengths. Grant awarded by the John Templeton Foundation (ID#47910).