# Validity and Reliability

# Validity

# Historical Views of Validity

- Early views saw validity as a static property
  - And an instrument as either valid or not
    - "By validity it is meant the degree to which a test or examination measures what it purports to measure." (Ruch, 1924)

# Historical Views of Validity

- Early views saw validity as a static property (cont.)
  - Often as evidenced with a correlation with an outside measure
    - E.g., Guilford (1946, p. 429):
      - "[I]n a very general sense, a test is valid for anything with which it correlates."

# Historical Views of Validity

- 1950s saw a seminal change to the field
  - With a broadening view of validity
- E.g., Campbell & Fiske (1959)
  - Argued for discrete types of validity
    - And needs for multiple kinds of evidence

# Historical Views of Validity

- Publication of *Standards for Educational and Psychological Testing* (APA, AERA, & NCME,1966)
  - Non-unitary, less static view of validity
    - An instrument is valid to the extent to which it produces information useful for a given purpose

# Historical Views of Validity

- Publication of *Standards* (APA, AERA, & NCME, 1966; cont.)
  - Included the "trinity" view of validity
    - First posited by Cronbach & Meehl (1955)
    - Viz.:
      - Construct validity
      - Content validity
      - Criterion validity

# Construct Validity

- The extent to which the instrument measures the intended (non-ostensible) construct
- Considered by some to subsume content & criterion validities

# Construct Validity (cont.)

- Distinguished from “face validity”
  - Construct validity typically requires content experts
  - Face validity can use lay views

# Content Validity

- "The extent to which a measure represents all facets of a given construct" (Wikipedia)
- I.e., measures the full range
  - And/or all dimensions of a multi-dimenstional trait

# Criterion Validity

- How well an instrument measures relevant outcomes
  - Do its measures correspond with other measures of the same trait
- Sometimes subdivided into
  - Concurrent validity: Coeval predictions
  - Predictive validity: A priori predictions

# Historical Views of Validity (redux)

- By 1980s, emphasis shifted
  - To the inferences and decisions made from a given instrument
  - 1985 *Standards*:
    - An instrument's validity is "the **appropriateness**, **meaningfulness**, and **usefulness**" of its measurements

# Historical Views of Validity

- Importantly, the 1985 *Standards* also:
  - Conceived of validity support as a dynamic & on-going process
    - "[T]he process of accumulating evidence to support" inferences made
  - Began to deprecate the trinity view
    - "[T]he use of category labels should not be taken to imply . . . distinct types of validity"

# Validity as a Unitary Construct

- By the 1990s, consensus grew that validity is a unitary construct
  - With multiple lines of evidence supporting it
    - "Although many kinds of evidence may be used, we do not have different kinds of validity" (Kane, 1994, p. 136)

# Validity in the 1999 *Standards*

- "The inference regarding specific uses of a test are validated, not the test itself."
- "Rigorous distinctions between the categories [of types of validity] are not possible."
- "An ideal validation includes several types of evidence, which span" the trinity (p. 9)

# Validity in the 1999 *Standards*

- Validity is now "the degree to which evidence and theory support the interpretation of test scores by proposed uses of tests."

# Types of Evidence

- One supports valid uses by giving types of evidence:
  1. Construct-related evidence
  2. Content-related evidence
  3. Criterion-related evidence
  4. Validity generalization
  5. Differential prediction (DIF in IRT)

# Types of Evidence (cont.)

- Construct-Related Evidence
  - Measure of the non-ostensible domain of interest
- Content-Related Evidence
  - Extent to which items sample well the domain

# Types of Evidence (cont.)

- Criterion-related evidence
  - "How accurately can criterion performance be predicted from scores?"
- Validity generalization
  - How well uses can be "transported" between situations & applications

# Types of Evidence (cont.)

- Differential prediction
  - That the instrument may operate differently among different populations
  - A rather new aspect
    - That is related to considerations of the *consequences* of testing . . .

# Sources of Evidence

- In addition, the 1999 *Standards* proffer different sources of evidence, based on:
  1. Test content
  2. Response processes
  3. Internal structure
  4. Relationships to other variables
  5. Consequences of testing

# Sources of Evidence

- In addition, the 1999 *Standards* proffer different sources of evidence, based on:
  1. Test content
  2. Response processes
  3. Internal structure
  4. Relationships to other variables
  5. Consequences of testing

# Sources of Evidence (cont.)

- Test Content
  - Typically assessed via logical analyses and experts' evaluations
  - Assessments of:
    - Sufficiency
    - Clarity
    - Relevance
    - Match between items & construct

# Sources of Evidence (cont.)

- Test Content (cont.)
  - Also reviews:
    - Potential bias (culture, age, etc.)
    - Construct-irrelevant variance
      - Measuring *more* than it is intended to
    - Construct under-representation
      - Measuring *less* than it is intended to

# Sources of Evidence (cont.)

- In addition, the 1999 *Standards* proffer different sources of evidence, based on:
  1. Test content
  2. Response processes
  3. Internal structure
  4. Relationships to other variables
  5. Consequences of testing

# Sources of Evidence (cont.)

- Response Processes
  - Fit of response type with construct
  - E.g.,
    - Inclusion of social desirability or lack of self-awareness in self-report
    - Inability / inaccuracy of judges, e.g., to measure internal states from observations

# Sources of Evidence (cont.)

- In addition, the 1999 *Standards* proffer different sources of evidence, based on:
  1. Test content
  2. Response processes
  3. Internal structure
  4. Relationships to other variables
  5. Consequences of testing

# Sources of Evidence (cont.)

- Internal Structure
  - Match between item response patterns and internal constructs
    - E.g., test of confirmatory factor analysis
      - Or also perhaps DIF
    - Arguably over-emphasized given ease of conducting CFAs

# Sources of Evidence (cont.)

- In addition, the 1999 *Standards* proffer different sources of evidence, based on:
  1. Test content
  2. Response processes
  3. Internal structure
  4. Relationships to other variables
  5. Consequences of testing

# Sources of Evidence (cont.)

- Relationships to Other Variables
  - Subsumes many "legacy" types of validity
    - E.g.,
      - Convergent & divergent validity
      - Comparisons of performance differences / similarities across groups
      - Studies of validity generalizations

# Sources of Evidence (cont.)

- In addition, the 1999 *Standards* proffer different sources of evidence, based on:
  1. Test content
  2. Response processes
  3. Internal structure
  4. Relationships to other variables
  5. Consequences of testing

# Sources of Evidence (cont.)

- Consequences of Testing
  - The positive & negative ramifications of being tested / given scores
  - Only briefly mentioned before the 1999 *Standards*
    - This remains the most controversial source of validity evidence
    - Being new, there are fewer guidelines for its assessment

# Summary of Types of Evidence

| 1985 "Trinity" Types | 1999 *Standards* Evidence based on: |
|---|---|
| Construct-related evidence (also subsumes content-related evidence) | Test Content |
| | Response Processes |
| | Internal Structure |
| Criterion-related evidence | Relationships to Other Variables |
| | Consequences of Testing |

| "Trinity" | '99 Sources: | Examples of Types of Evidence |
|---|---|---|
| Construct-related (and content-related) evidence | Test Content | Logical analyses & experts reviews of representativeness of items to domain, extent items span domain, clarity items, construct irrelevance, under-representation; extent any of these introduce bias |
| | Response Processes | Respondent interviews; studies of response patterns across populations; studies of how judges, researchers, etc. collect & interpret responses |
| | Internal Structure | Factor- and cluster-analytic studies; item analyses of inter-relationships; differential item functioning (DIF) via item response theory (IRT) |
| | Relationships to Other Variables | Convergence & discrimination studies (e.g., multi-trait &-method studies, p. 231); Hypothesis tests of effects of interventions on test scores; Known-group comparison & longitudinal studies studies on expected outcomes |
| Criterion-related evidence | | Correlations of scores with external, criterion variables measuring strength, directionality of relationships; Theory-guided group separation studies testing predictiveness of scores on relevant outcomes across & between populations; Differential group relationships and prediction studies; Studies of effectiveness of selections, classifications, & placements; Validity generalization studies |
| | Consequences of Testing | Studies of the extent to which expected/anticipated benefits or unexpected/unanticipated consequences are realized |

# Does This Matter?

- The current view does represent a more sophisticated perspective
  - That addresses how validity is actually used by the field
- But, the field has been slow to adopt it
  - So, your adoption of it may be warranted but under-appreciated

# Reliability

# Classical Measurement Theory

- CMT models observed measurements ($O$) as composed of

  $$O = T + E$$

  - $T$ = True scores
    - Fixed for any given point in time
  - $E$ = Error
    - Unrelated to one's true score ($r_{TE} = 0$)
    - With mean = 0
    - Normally-distributed variance

# Classical Definition of Reliability

- Within a sample of measurements:

$$Var\,(O) = Var\,(T) + Var\,(E)$$

- Standardizing on observed scores:

$$\frac{Var(O)}{Var(O)} = \frac{Var(T)}{Var(O)} + \frac{Var(E)}{Var(O)} = 1$$

# Classical Def. of Reliability (cont.)

- Classical definition of reliability:

$$\frac{Var(T)}{Var(O)}$$

  - This *variance ratio* is equivalent to a squared correlation
  - Reliability, then, is $r_{TO}^2$
    - Denoted the reliability coefficient

# Classical Def. of Reliability (cont.)

- And since Var ($O$) = Var ($T$) + Var ($E$):

$$\frac{Var(T)}{Var(T)+Var(E)}=\frac{Signal}{Signal+Noise}$$

# Example

| Respondent | Observed Score ($X_o$) | | True Score ($X_t$) | | Error ($X_e$) |
|---|---|---|---|---|---|
| Ashley | 120 | = | 130 | + | -10 |
| Bob | 145 | = | 120 | + | 25 |
| Carl | 95 | = | 110 | + | -15 |
| Denise | 85 | = | 100 | + | -15 |
| Eric | 115 | = | 90 | + | 25 |
| Felicia | 70 | = | 80 | + | -10 |
| Mean | 105.00 | | 105 | | 0 |
| Variance | 608.33 | | 291.67 | | 316.67 |
| Std. Dev. | 24.66 | | 17.08 | | 17.80 |

# Example (cont.)

| Respondent | ($X_o$) Observed Score | | ($X_t$) True Score | | ($X_e$) Error |
|---|---|---|---|---|---|
| Ashley | 120 | = | 130 | + | -10 |
| Bob | 145 | = | 120 | + | 25 |
| Carl | 95 | = | 110 | + | -15 |
| Denise | 85 | = | 100 | + | -15 |
| Eric | 115 | = | 90 | + | 25 |
| Felicia | 70 | = | 80 | + | -10 |
| Mean | 105.00 | | 105 | | 0 |
| Variance | 608.33 | | 291.67 | | 316.67 |
| Std. Dev. | 24.66 | | 17.08 | | 17.80 |

$$\bar{X}_E = 0$$

# Example

| Respondent | $(X_o)$ Observed Score | | $(X_t)$ True Score | | $(X_e)$ Error |
|---|---|---|---|---|---|
| Ashley | 120 | = | 130 | + | -10 |
| Bob | 145 | = | 120 | + | 25 |
| Carl | 95 | = | 110 | + | -15 |
| Denise | 85 | = | 100 | + | -15 |
| Eric | 115 | = | 90 | + | 25 |
| Felicia | 70 | = | 80 | + | -10 |
| Mean | 105.00 | | 105 | | 0 |
| Variance | 608.33 | | 291.67 | | 316.67 |
| Std. Dev. | 24.66 | | 17.08 | | 17.80 |

$$\bar{X}_E = 0$$

$$r_{TE} = 0$$

# Example (cont.)

- Since $Var = S^2$:

$$S_O^2 = S_T^2 + S_E^2$$

# Example (cont.)

| Respondent | ($X_o$) Observed Score | | ($X_t$) True Score | | ($X_e$) Error |
|---|---|---|---|---|---|
| Ashley | 120 | = | 130 | + | -10 |
| Bob | 145 | = | 120 | + | 25 |
| Carl | 95 | = | 110 | + | -15 |
| Denise | 85 | = | 100 | + | -15 |
| Eric | 115 | = | 90 | + | 25 |
| Felicia | 70 | = | 80 | + | -10 |
| Mean | 105.00 | | 105 | | 0 |
| Variance | 608.33 | | 291.67 | | 316.67 |
| Std. Dev. | 24.66 | | 17.08 | | 17.80 |

$$\begin{array}{r} 291.67 \\ +\ 316.67 \\ \hline 608.33 \end{array}$$

# Conceptualizing Reliability

- Mathematically identical
  - I.e., identical values for the coefficient of reliability, $R_{xx}$
  - However, they
    - emphasize different facets of reliability's meaning
    - are all common ways of discussing reliability

# Conceptualizing Reliability (cont.)

| Statistical Basis of Reliability, in terms of: | Conceptual Basis of Reliability: Observed score in relation to: | |
|---|---|---|
| | True Scores | Measurement Error |
| Proportions of Variance | Ratio of true score variance to observed score variance | Lack of error variance |
| Correlations | (Squared) correlation between observed & true scores | Lack of correlation btn observed & true |

# Conceptualizing Reliability (cont.)

| Statistical Basis of Reliability, in terms of: | Conceptual Basis of Reliability: Observed score in relation to: | |
|---|---|---|
| | True Scores | Measurement Error |
| Proportions of Variance | Ratio of true score variance to observed score variance $R_{XX} = \frac{S_T^2}{S_O^2}$ | Lack of error variance |
| Correlations | (Squared) correlation between observed & true scores | Lack of correlation btn observed & true |

# Example (True Score)

| Respondent | ($X_o$) Observed Score | | ($X_t$) True Score | | ($X_e$) Error |
|---|---|---|---|---|---|
| Ashley | 120 | = | 130 | + | -10 |
| Bob | 145 | = | 120 | + | 25 |
| Carl | 95 | = | 110 | + | -15 |
| Denise | 85 | = | 100 | + | -15 |
| Eric | 115 | = | 90 | + | 25 |
| Felicia | 70 | = | 80 | + | -10 |
| Mean | 105.00 | | 105 | | 0 |
| Variance | 608.33 | | 291.67 | | 316.67 |
| Std. Dev. | 24.66 | | 17.08 | | 17.80 |

$$R_{XX} = \frac{S_T^2}{S_O^2}$$

$$R_{XX} = \frac{291.67}{608.33}$$

$$R_{XX} = .48$$

# Conceptualizing Reliability (cont.)

| Statistical Basis of Reliability, in terms of: | Conceptual Basis of Reliability: Observed score in relation to: | |
|---|---|---|
| | True Scores | Measurement Error |
| Proportions of Variance | Ratio of true score variance to observed score variance $R_{XX} = \frac{S_T^2}{S_O^2}$ | Lack of error variance $R_{XX} = 1 - \frac{S_E^2}{S_O^2}$ |
| Correlations | (Squared) correlation between observed & true scores | Lack of correlation btn observed & true |

# Example (Measurement Error)

| Respondent | ($X_o$) Observed Score | | ($X_t$) True Score | | ($X_e$) Error |
|---|---|---|---|---|---|
| Ashley | 120 | = | 130 | + | -10 |
| Bob | 145 | = | 120 | + | 25 |
| Carl | 95 | = | 110 | + | -15 |
| Denise | 85 | = | 100 | + | -15 |
| Eric | 115 | = | 90 | + | 25 |
| Felicia | 70 | = | 80 | + | -10 |
| Mean | 105.00 | | 105 | | 0 |
| Variance | 608.33 | | 291.67 | | 316.67 |
| Std. Dev. | 24.66 | | 17.08 | | 17.80 |

$$R_{XX} = 1 - \frac{S_E^2}{S_O^2}$$

$$R_{XX} = 1 - \frac{316.67}{608.33}$$

$$R_{XX} = .48$$

# Conceptualizing Reliability (cont.)

| Statistical Basis of Reliability, in terms of: | Conceptual Basis of Reliability: Observed score in relation to: | |
|---|---|---|
| | True Scores | Measurement Error |
| Proportions of Variance | Ratio of true score variance to observed score variance $R_{XX} = \frac{S_T^2}{S_O^2}$ | Lack of error variance $R_{XX} = 1 - \frac{S_E^2}{S_O^2}$ |
| Correlations | (Squared) correlation between observed & true scores $R_{XX} = r_{TO}^2$ | Lack of correlation btn observed & true |

# Example (Measurement Error)

$r_{TO} = .69$

| Respondent | ($X_o$) Observed Score | | ($X_t$) True Score | | ($X_e$) Error |
|---|---|---|---|---|---|
| Ashley | 120 | = | 130 | + | -10 |
| Bob | 145 | = | 120 | + | 25 |
| Carl | 95 | = | 110 | + | -15 |
| Denise | 85 | = | 100 | + | -15 |
| Eric | 115 | = | 90 | + | 25 |
| Felicia | 70 | = | 80 | + | -10 |
| Mean | 105.00 | | 105 | | 0 |
| Variance | 608.33 | | 291.67 | | 316.67 |
| Std. Dev. | 24.66 | | 17.08 | | 17.80 |

$$R_{XX} = r_{TO}^2$$

$$R_{XX} = .69^2$$

$$R_{XX} = .48$$

# Conceptualizing Reliability (cont.)

| Statistical Basis of Reliability, in terms of: | Conceptual Basis of Reliability: Observed score in relation to: | |
|---|---|---|
| | True Scores | Measurement Error |
| Proportions of Variance | Ratio of true score variance to observed score variance $R_{XX} = \frac{S_T^2}{S_O^2}$ | Lack of error variance $R_{XX} = 1 - \frac{S_E^2}{S_O^2}$ |
| Correlations | (Squared) correlation between observed & true scores $R_{XX} = r_{TO}^2$ | Lack of correlation btn observed & true $R_{XX} = 1 - r_{EO}^2$ |

# Example (Measurement Error)

$$r_{TO} = .72$$

| Respondent | ($X_o$) Observed Score | | ($X_t$) True Score | | ($X_e$) Error |
|---|---|---|---|---|---|
| Ashley | 120 | = | 130 | + | -10 |
| Bob | 145 | = | 120 | + | 25 |
| Carl | 95 | = | 110 | + | -15 |
| Denise | 85 | = | 100 | + | -15 |
| Eric | 115 | = | 90 | + | 25 |
| Felicia | 70 | = | 80 | + | -10 |
| Mean | 105.00 | | 105 | | 0 |
| Variance | 608.33 | | 291.67 | | 316.67 |
| Std. Dev. | 24.66 | | 17.08 | | 17.80 |

$$R_{XX} = 1 - r_{EO}^2$$

$$R_{XX} = 1 - .72^2$$

$$R_{XX} = 1 - .52$$

$$R_{XX} = .48$$

# Example (cont.)

- Remember $R_{XX}$ is a variance ratio
  - Which is equivalent to a squared correlation
  - Since $R_{XX}$ = .48,
    - .48 (48%) of the variance here is attributable to true scores

# True Score = Domain Score

- Items sample a domain
  - Like other samples, they only estimate the population
- True score would be one's scores on *all* items in that domain
  - Thus reliability tests attempt to measure how well items represent the domain

# Actual Reliability Tests

- Four primary estimates
  1. Internal consistency
  2. Inter-rater
  3. Intra-rater / Test-retest
  4. Parallel form

# Actual Reliability Tests

- Four primary estimates
  1. Internal consistency
     - Correlation between **items**
  2. Inter-rater
     - Correlation between **raters**
  3. Intra-rater / Test-retest
     - Correlation between **administrations**
  4. Parallel form
     - Correlation between **versions**

# Internal Consistency

- Common measures of internal consistency
  - Coefficient α
    - Used for interval / ratio data
    - May underestimate associations in ordinal, so ordinal α is better
  - Kuder-Richardson Formulae 20 & 21
    - Used for dichotomous data

# Coefficient α

- aka Cronbach's α
- Conceptually
  - How well any item score predicts any other item score
  - Or the mean of the distribution of all split-half correlations
    - Thus better than split-half tests

# Coefficient α (cont.)

- Also conceptually:

$$\alpha = \frac{N \times \bar{c}}{\left(\bar{v} + \left(N - 1\right) \times \bar{c}\right)}$$

  - $\bar{N}$ = number of items
  - $\bar{c}$ = mean covariance between item pairs
  - $v$ = mean item variance

# Coefficient α (cont.)

- Generally acceptable levels
  - Excellent $\alpha \geq .9$
  - Good $.9 > \alpha \geq .8$
  - Acceptable $.8 > \alpha \geq .7$
  - Questionable $.7 > \alpha \geq .6$
  - Poor $.6 > \alpha \geq .5$
  - Unacceptable $.5 > \alpha$

# Coefficient α (cont.)

- Coefficient α considerations
  - Sensitive to number of items:

$$\alpha = \frac{N \times \bar{c}}{\left(\bar{v} + \left(N-1\right) \times \bar{c}\right)}$$

    - Adding relevant items can increase it
    - But very high levels may imply redundant items

# Coefficient α (cont.)

- Coefficient α considerations (cont.)
  - Also sensitive to total variance
    - Adding non-redundant items from same domain can increase it
    - Sampling a heterogeneous group of participants can also increase it

# Coefficient α (cont.)

- Coefficient α considerations (cont.)
  - Low levels may imply:
    - Nonunitary instrument
    - Skewed distribution of scores
  - High (> 15%) rates of missing data can inflate α
    - Especially if missingness is non-random

# KR 20 / KR 21

- Both are measures of consistency of results
- KR 20
  - Used for items of varying difficulty
- KR 21
  - Used for items of equal difficulty

# KR 20 / KR 21 (cont.)

$$KR20 = \frac{N}{N-1} \times \frac{\sum \rho q}{Var}$$

- $N$ = number of items
- $\rho$ = proportion of participants "passing"
- $q$ = proportion of participants "failing"
- $Var$ = Total test variance

# KR 20 / KR 21 (cont.)

$$KR21 = \frac{n}{n-1} \times \frac{M(n-M)}{n \times Var}$$

- $n$ = number of participants
- $M$ = mean score on test
- $Var$ = Total test variance

# KR 20 / KR 21 (cont.)

- KR 20 / 21 are also sensitive to:
  - Instrument length
    - But less than coefficient $\alpha$
  - Total instrument variance
  - Missing data
    - Especially since $q$ can also include missing as well as "fails"

# Actual Reliability Tests

- Four primary estimates

1. Internal consistency
2. **Inter-rater**
3. Intra-rater / Test-retest
4. Parallel form

# Interrater Reliability

- Agreement is between raters, not items
- For 2 raters:
  - Nominal: $\chi^2$ (or Cramer's *V*, etc.)
  - Ordinal: Spearman's $\rho$ (or Kendall's $\tau$)
  - Interval: Pearson's *r*
- For >2 raters, use coefficient $\alpha$

# Actual Reliability Tests

- Four primary estimates

1. Internal consistency
2. Interrater
3. **Intrarater / Test-retest**
4. Parallel form

# Intrarater & Test-Retest

- Typically uses Pearson's *r*
  - Like test-retest, we strive for independent scores at each wave
    - Waltz et al. (2017) recommend ~2 weeks
      - And to shuffle items
    - Ensure similar administration conditions
  - Interested here in correlation, not matching scores per se

# Intrarater & Test-Retest (cont.)

- If indeed interested in matching scores
  - Compute percentage of agreement
    - I.e., percent of times rater(s) assign the same score to each item
    - Can be quite stringent for interval / ratio items
      - Also affected by test length (mean regression)

# Actual Reliability Tests

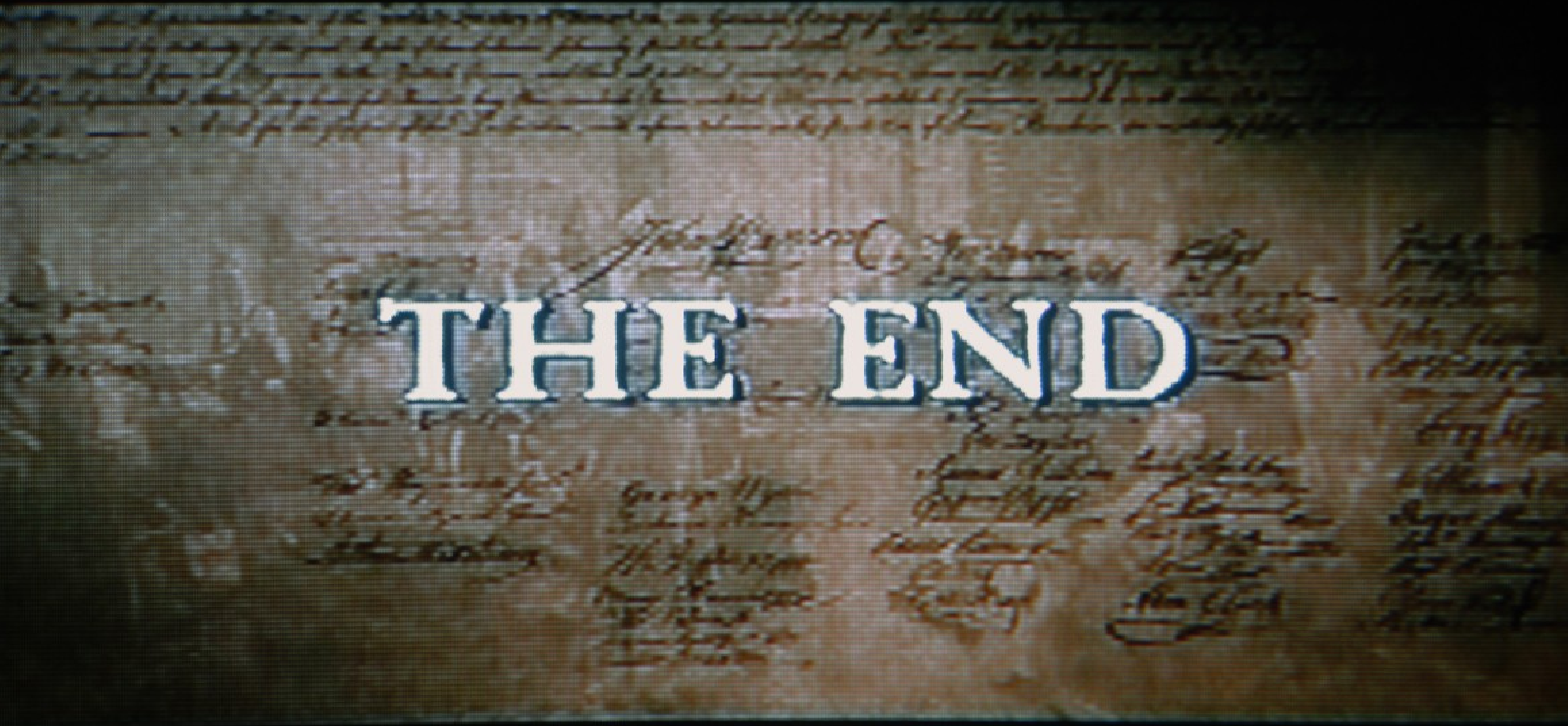
- Four primary estimates
  1. Internal consistency
  2. Inter-rater
  3. Intra-rater / Test-retest
  4. **Parallel form**

# Parallel Forms

- Correlation between scores on 2+ versions of an instrument
  - Instruments must be created as separate forms
    - I.e., not just split-half tests of an instrument

# Parallel Forms (cont.)

- Typically follows strong criterion-related evidence of both instruments' validity
  - E.g., first administering forms to same participants at same time
    - Testing means, variances, and convergence / discrimination with other relevant measures


THE END

# References


- APA, AERA, & NCME (1999). *Standards for Educational and Psychological Testing.*
- Campbell, D. T. & Fiske, D. W. (1959). Convergent and discrimination validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81.
- Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.
- Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*, 427–438.
- Kane, M. T. (1994). Validating interpretative arguments for licensure and certification examinations. *Evaluation & the Health Professions*, *17*, 133–159.
- Ruch, G. M. (1924). *The improvement of the written examination.* Chicago: Scott, Foreman and Company.