Lesson 3: Validity

Patrícia Martinková

Department of Statistical Modelling
Institute of Computer Science, Czech Academy of Sciences

Institute for Research and Development of Education
Faculty of Education, Charles University, Prague

NMST570, October 16, 2018

Table of contents

1. Review - Reliability
2. Validity
3. Test validation studies
4. Further issues
5. Conclusion

Review – Reliability

- Latent variable
- Measurement error
- Reliability
- Test-retest reliability
- Alternate forms
- Internal consistency
    - Split-half (first-second half, even-odd, random, average)
    - Cronbach's alpha
    - Kuder-Richardson formula
- Inter-rater reliability

Review – Reliability

- The degree to which an assessment tool produces stable and consistent results.
- Assuming $X = T + e$, true score and error uncorrelated
- Defined as squared correlation of the true and observed score $\mathrm{Rel}\,(X) = \rho_X = \mathrm{cor}^2(T, X) = \rho^2_{T,X}$
- Equivalently: the ratio of the true score variance to total observed variance $\rho_X = \frac{\mathrm{var}\,(T)}{\mathrm{var}\,(X)} = \frac{\sigma^2_T}{\sigma^2_X} = \frac{\sigma^2_X - \sigma^2_e}{\sigma^2_X} = 1 - \frac{\sigma^2_e}{\sigma^2_X}$
- Correlation between two independent, equaly precise measurements, measuring the same construct $\rho_X = \rho_{X_1, X_2}$
- $\rho_X \in \langle 0, 1\rangle$

Review – Reliability of composite measurements

- Goal is to provide multiple converging pieces of information
- E.g. educational tests, scales, questionnaires, ...

What is the relationship between reliability of composite measurement $X = \sum_{j=1}^{m} X_j$ and reliability of its components?

Spearman-Brown prophecy formula (1910)

Assume m parallel measurements $X_1, \ldots, X_m$ (independent, equally precise, with uncorrelated errors and uncorrelated with true scores). Then reliability of each X_i is the same ρ and the reliability of composite measurement X is

$$\rho_X = \frac{m \cdot \rho}{1 + (m-1)\rho}$$

Remark: Adding parallel items increases reliability of total score.

Generalized prophecy formula

Spearman-Brown prophecy formula (generalized)

Assume test composed of m_1 parallel measurements $X = \sum_{j=1}^{m_1} X_j$ and its prolonged or shortened version composed of m_2 parallel measurements $X = \sum_{j=1}^{m_2} X_j$. Then the relationship between their reliabilities is

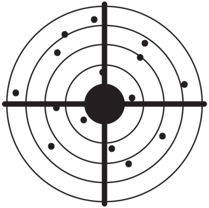
$$\rho_{m_2} = \frac{\frac{m_2}{m_1} \cdot \rho_{m_1}}{1 + (\frac{m_2}{m_1} - 1)\rho_{m_1}}$$

Proof (hint): Notice that

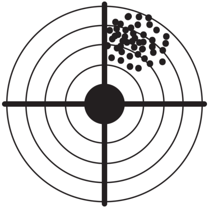
$$\rho_1 = \frac{\frac{1}{m_1} \cdot \rho_{m_1}}{1 + (\frac{1}{m_1} - 1)\rho_{m_1}} = \frac{\frac{1}{m_2} \cdot \rho_{m_2}}{1 + (\frac{1}{m_2} - 1)\rho_{m_2}}$$

Review - Reliability and Validity

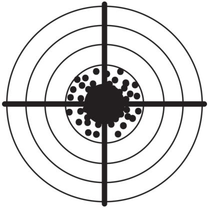
- high reliability does not ensure high validity
- validity is bounded by reliability



Low reliability thus low validity



High reliability but low validity



High reliability and high validity

$$\mathrm{cor}\left(X_1, X_2\right) = \frac{\mathrm{cov}\left(X_1, X_2\right)}{\sqrt{\mathrm{var}\left(X_1\right)}\sqrt{\mathrm{var}\left(X_2\right)}} = \frac{\mathrm{cov}\left(T_1, T_2\right) + 0 + 0 + 0}{\sqrt{\mathrm{var}\left(T_1\right)\frac{\mathrm{var}\left(X_1\right)}{\mathrm{var}\left(T_1\right)}}\sqrt{\mathrm{var}\left(T_2\right)\frac{\mathrm{var}\left(X_2\right)}{\mathrm{var}\left(T_2\right)}}}$$

$$= \mathrm{cor}\left(T_1, T_2\right)\sqrt{\mathrm{Rel}\left(X_1\right)\mathrm{Rel}\left(X_2\right)}$$

Test validity

- The degree to which evidence and theory support the interpretations of test scores
- The degree to which test measures what it is supposed to measure

- Content-related
  - Face validity
  - Construct validity
  - Content validity

- Criterion-related
  - Concurrent
  - Predictive
  - Incremental

Content-related validity

Construct validity

- Extent to which a test captures a specific theoretical construct
- Subsumes other types of validity
  - convergent validity: associated with things it should be
  - discriminant validity: not associated with things it should not be
- Needs empirical and theoretical evidence
  - analyses of the internal structure (correlations between item answers)

Content validity

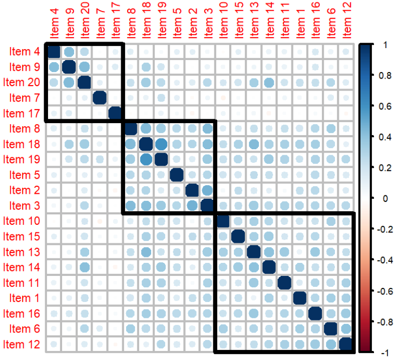
- Does the test cover the domain to be measured?
- Needs careful selection of which items to include

Face validity

- Does the test “apear to” measure what it aims to?
  (to a member of target population)
  - Advantage: respondent can use context to help interpret the question
  - Disadvantage: respondent might try to "bend & shape"their answers

Analysis of internal structure

- Correlation between answers to individual items
- Factor analysis
- Cluster analysis



Criterion-related validity

Concurrent validity

- Correlation with other measures of the same construct that are measured at the same time
- E.g. admission test and IQ test

Predictive validity

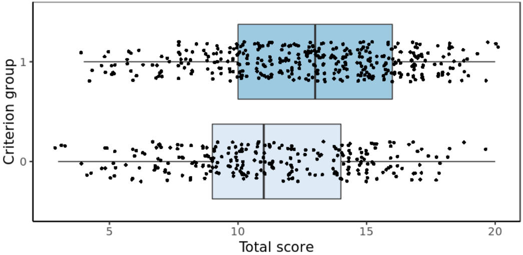
- Correlation with other measures of the same construct that are measured later
- E.g. admission test and subsequent GPA or study success

Incremental validity

- Increase of predictive validity, adds information beyond that provided by an existing methods
- Usually assessed by multiple regression
- E.g. admission test adds to prediction of subsequent GPA above high-school GPA

Criterion-related validity

- Correlation
- Regression
  - Linear, logistic
  - Multiple (accounting for more characteristics)
  - Hierarchical (accounting for hierarchical structure - countries/schools/classes)



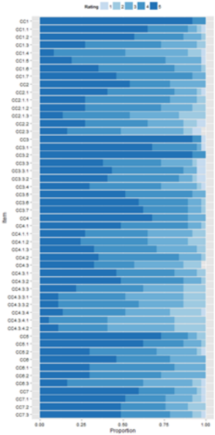
Example 1: Physiology concept inventories

http://www.physiologyconcepts.org/

Biology Education Research Group (BERG, University of Washington)

Cell-cell communication (CCC) conceptual framework

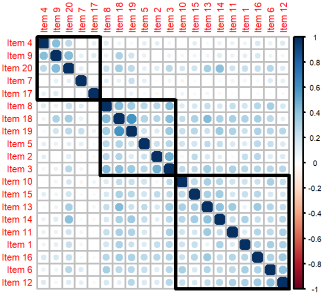
- Study develops and validates hierarchical CCC conceptual framework
- Validation based on responses of undergraduate biology faculty
- Subsequently can be used for development and construct validation of related test on CCC



Michael J, Martinková P, McFarland JL, Wright A, Cliff W, Modell H, Wenderoth MP. Validating a conceptual framework for the core concept of "cell-cell communications". Advances in Physiology Education, Vol. 41 no. 2, pp. 260-265, 2017.

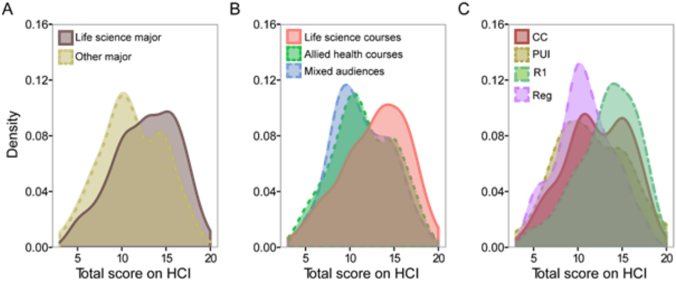
Homeostasis concept inventory (HCI)

- HCI developed based on Homeostasis conceptual framework (HCF)
- Items test knowledge of individual elements in HCF



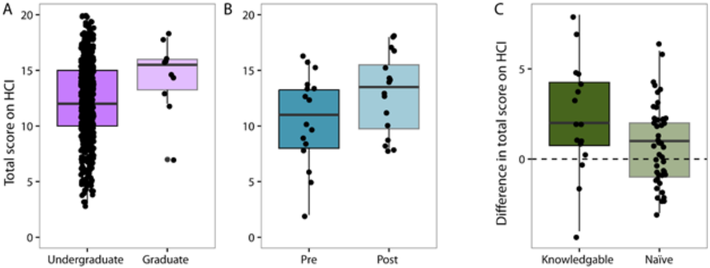
McFarland, Price, Wenderoth, Martinková, et al. Development and Validation of the Homeostasis Concept Inventory. CBE Life Sciences Education, 16(2), ar35, 2017.

Homeostasis concept inventory (HCI)



McFarland, Price, Wenderoth, Martinková, et al. Development and Validation of the Homeostasis Concept Inventory. CBE Life Sciences Education, 16(2), ar35, 2017.

Homeostasis concept inventory (HCI)



McFarland, Price, Wenderoth, Martinková, et al. Development and Validation of the Homeostasis Concept Inventory. CBE Life Sciences Education, 16(2), ar35, 2017.

Example 2: Assessment set for Multiple Sclerosis

- Set of 11 clinical tests with 2-20 items/components
- Reliability
  - Internal consistency (Cronbach's alpha)
  - Test-retest
  - Validity
- Validity
  - Stability without treatment
  - Changes after treatment
  - Correlations with EDSS
  - Correlations between individual clinical tests

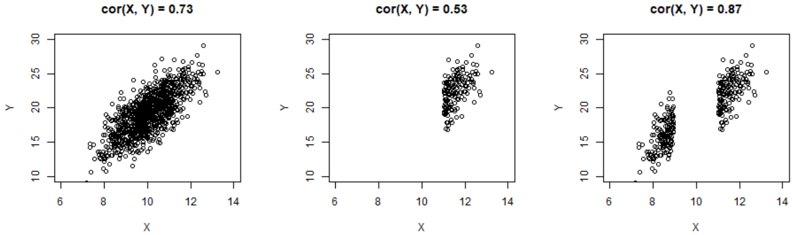
Řasová K, Martinková P, Vyskotová J, Šedová M. Assessment set for evaluation of clinical outcomes in multiple sclerosis - psychometric properties. Patient Related Outcome Measures, 3, pp. 59-70, 2012.

Further issues in validity and reliability

- Restriction in range
    - Correction of validity estimate
    - Correction of reliability estimate
- Effect of unreliability on validity

Restriction of range

- Common problem in validation studies
- Restriction in range of the predictor variable
- Example:
  - Observing only admitted students
  - Observing only students who did not pass the first exam

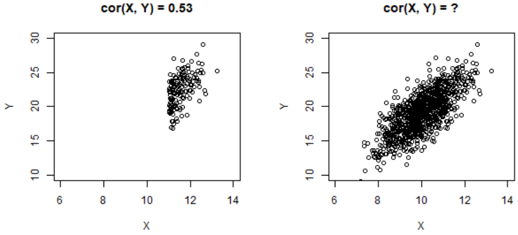


cor(X, Y) = 0.73

cor(X, Y) = 0.53

cor(X, Y) = 0.87

Correction for range restriction

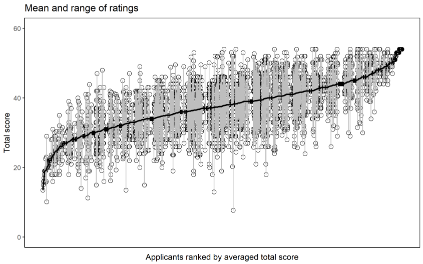


Correction for range restriction (see e.g. Wiberg & Sundstrom, 2009)

$$r_{XY} = \frac{s_X r_{x,y}}{\sqrt{s_X^2 r_{x,y}^2 + s_x^2 - s_x^2 r_{x,y}^2}} = \frac{0.99 \cdot 0.53}{\sqrt{0.99^2 \cdot 0.53^2 + 0.46^2 + 0.46^2 \cdot 0.53^2}} = 0.80$$

$r_{xy} = 0.53$ observed correlation btw X and Y in restricted sample
$s_x = 0.46$ estimated SD of X in restricted sample
$s_X = 0.99$ estimated SD of X in original sample

Reliability – Restriction to range

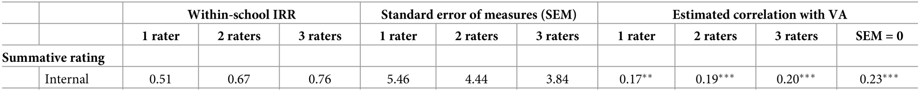


- Similar issue for reliability estimate
- Having restricted sample, estimate of reliability may be unproper
- Correction to restriction of range (see e.g. Fife et al., 2012)

$$\rho_X = 1 - \frac{\sigma_x^2}{\sigma_X^2}(1 - \rho_x)$$

Correction for unreliability

Example: Ratings of teacher applicants



		Within-school IRR			Standard error of measures (SEM)			Estimated correlation with VA			
		1 rater	2 raters	3 raters	1 rater	2 raters	3 raters	1 rater	2 raters	3 raters	SEM = 0
Summative rating											
	Internal	0.51	0.67	0.76	5.46	4.44	3.84	0.17^{**}	0.19^{***}	0.20^{***}	0.23^{***}

- (SB) prophecy formula to estimate reliability of average rating:

$$IRR_{\bar{Y}} = \frac{\sigma_A^2 + \sigma_S^2 + \sigma_{AS}^2}{\sigma_A^2 + \sigma_S^2 + \sigma_{AS}^2 + \sigma_B^2/J + \sigma_e^2/J}$$

- Standard error of measures: $\sigma_B^2/J + \sigma_e^2/J$
- Attenuation formula to estimate corrected correlation with VA:

$$\mathrm{cor}\left(\bar{Y}, \mathrm{VA}\right) = \mathrm{cor}\left(T, \mathrm{VA}\right)\sqrt{IRR_{\bar{Y}}}$$

Martinková et al (2018). Disparities in ratings of internal and external applicants...

https://doi.org/10.1371/journal.pone.0203002

Conclusion

In this presentation, we have

- Presented most important aspects/types of test validity
    - Content-related
        - Construct validity
        - Content validty
        - Face validity
    - Criterion-related
        - Concurrent validity
        - Predictive validity
        - Incremental validity
- Presented examples of test validation studies
- Presented further issues in validity estimation
    - Correction for range restriction
    - Correction for unreliability

Thank you for your attention!

www.cs.cas.cz/martinkova

References

- Kane, MT (2006). Validation. In Brennan RL (ed.), (2006). *Educational Measurement (4th edn.).* Westport, CT: Praeger. pp. 65–110.
- Michael J, Martinková P, McFarland JL, Wright A, Cliff W, Modell H, Wenderoth MP. Validating a conceptual framework for the core concept of "cell-cell communications". Advances in Physiology Education, Vol. 41 no. 2, pp. 260-265, 2017.
- McFarland, Price, Wenderoth, Martinková, et al. (2017). Development and Validation of the Homeostasis Concept Inventory. CBE Life Sciences Education, 16(2), ar35. doi 10.1187/cbe.16-10-0305
- Řasová K, Martinková P, Vyskotová J, Šedová M. Assessment set for evaluation of clinical outcomes in multiple sclerosis - psychometric properties. Patient Related Outcome Measures, 3, pp. 59-70, 2012.
- Martinková P, Goldhaber D, & Erosheva E. (2018). Disparities in ratings of internal and external applicants: A case for model-based inter-rater reliability. *PLOS ONE*, 13(10): e0203002. https://doi.org/10.1371/journal.pone.0203002

Vocabulary

- Validity
  - Content-related
    - Construct validity
    - Content validty
    - Face validity
  - Criterion-related
    - Concurrent validity
    - Predictive validity
    - Incremental validity
- Correction for range restriction
- Correction for unreliability