

Measurement Matters Resource List

[Eiko Fried](#) & [Jessica Flake](#)

Last updated on November 30 2019

This resource list contains reading material on the topic of measurement in psychological sciences. We hope the list will be a useful tool in helping researchers to improve measurement practices, and inspire debates about measurement in psychology. We initiated the repository originally as accompanying material to our piece in the APS Observer entitled “[Measurement Matters](#)”. We consider it to be a preliminary, active, living document, and plan to update it regularly. We also want to acknowledge that the list is the outcome of many different sources, such as the SIPS pre-conference at SPSP 2018. If you have other papers you would like to see included here, please let us know (eikofried@gmail.com & kayflake@gmail.com).

This is not a complete overview of all relevant papers on measurement in psychology, but a selection of useful papers. We don't agree with all positions put forward, but believe the papers and books provide a healthy balance of viewpoints. We intend this list as a resource for researchers at all levels of measurement expertise, and marked a few papers with * that we consider to be exceptionally suitable introductory papers for beginners. You can find the list on the Open Science Framework at <https://osf.io/zrkd4>.

The document is structured as follows:

1. Validity theory and philosophy of measurement in psychology
 - 1.1. Foundations
 - 1.2. Overviews & syntheses
 - 1.3. Critical and Philosophical Discussions on Measurement & Validity Theory
2. Key Topics in Measurement
 - 2.1. Theory and Item Development
 - 2.2. Formative vs. Reflective Measurement
 - 2.3. Reliability
 - 2.4. Measurement of Individual Differences
3. Quantitative Approaches and Models
 - 3.1. Factor Analysis
 - 3.2. Measurement Invariance
 - 3.3. Item Response Theory and Differential Item Functioning
 - 3.4. Mixture Models (Latent Class/Profile Analysis)
 - 3.5. Network Models
4. Analyses of Measurement Practices in Psychology
5. Examples of Construct Validation Efforts in Substantive Research
6. Reference Texts and Book Chapters

1. Validity Theory and Philosophy of Measurement in Psychology

This section introduces major approaches to the theory and philosophy of measurement in psychology, and is structured into 3 parts: seminal works that introduce big ideas and frameworks (i.e., foundational readings), useful overviews/syntheses, and critical texts that address philosophical issues in validity theory.

1.1 Foundations

Borsboom, D., & Mellenbergh, G. J. (2004). The Concept of Validity. *Psychological Review*, 111(4), 1061–1071. <http://doi.org/10.1037/0033-295X.111.4.1061>

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2), 81.

* Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3(3), 380.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), 1–73. <http://doi.org/10.1111/jedm.12000>

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–694.

1.2 Overviews and Syntheses

* Benson, J. (1998). Developing a Strong Program of Construct Validation: A Test Anxiety Example. *Educational Measurement: Issues and Practice*, 17(1), 10–17.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110(2), 305–314.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, (451), 425–440. <http://link.springer.com/article/10.1007/s11336-006-1447-6>

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309–319. <http://doi.org/10.1037/1040-3590.7.3.309>

* Edwards, J. R., & Bagozzi, R. P. (2000). On the Nature and Direction of Relationships Between Constructs and Measures. *Psychological Methods*, 5(2). <http://doi.org/10.1037/1082-989X.5.2>
Mcgrath, R. E. (2005). Conceptual complexity and construct validity. *Journal of Personality Assessment*, 85(2), 37–41. <http://doi.org/10.1207/s15327752jpa8502>

Slaney, K. (2017). Validating psychological constructs: Historical, philosophical, and practical dimensions. (J. Martin, Ed.). London: Palgrave Macmillian.
<http://doi.org/10.1057/978-1-137-38523-9>

Strauss, M. E., & Smith, G. T. (2009). Construct validity: advances in theory and methodology. *Annual Review of Clinical Psychology*, 5, 1–25. <http://doi.org/10.1146/annurev.clinpsy.032408.153639>

1.3 Critical and Philosophical Discussions of Measurement & Validity Theory

Borsboom, D., Rhemtulla, M., Cramer, A. O. J., van der Maas, H. L. J., Scheffer, M., & Dolan, C. V. (2016). Kinds versus continua: a review of psychometric approaches to uncover the structure of psychiatric constructs. *Psychological Medicine*, 1–13.
<http://doi.org/10.1017/S0033291715001944>

Fried, E. I. (2017). What are psychological constructs? On the nature and statistical modeling of emotions, intelligence, personality traits and mental disorders. *Health Psychology Review*, 11(2), 130–134. <http://doi.org/10.1080/17437199.2017.1306718>

Kendler, K. S. (2016). The nature of psychiatric disorders. *World Psychiatry*, 15(1), 5–12. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/26833596>

Maul, A. (2017). Rethinking Traditional Methods of Survey Validation. *Measurement*, 15(2), 51–69.
<http://doi.org/10.1080/15366367.2017.1348108>

* Meehl, P. E. (1978). Theoretical risks and tabular asterisks: The slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4), 806–834.
<http://doi.org/10.1037//0022-006X.46.4.806>

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88(3), 355–383. <http://doi.org/10.1111/j.2044-8295.1997.tb02641.x>

2. Key Topics in Measurement

This section includes papers that focus on key topics and approaches related to the measurement of psychological constructs. We have included didactic introductions as well as tutorials in this section. While measurement and modeling can be hard to disentangle, the present section puts the focus more on the measurement side, whereas the section 3. is focused on modeling.

2.1 Theory and Item Development

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology*, 34(4), 481.

* Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, 15(4), 380–387. <http://doi.org/10.1037/a0025704>

Simms, L. J. (2008). Classical and Modern Methods of Psychological Scale Construction. *Social and Personality Psychology Compass*, 2(1), 414–433. <http://doi.org/10.1111/j.1751-9004.2007.00044.x>

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of applied psychology*, 47(2), 149.

2.2 Measurement: Formative vs. Reflective Models

Bollen, K. A., Diamantopoulos, A., & Bollen, K. A. (2015). Psychological Methods In Defense of Causal – Formative Indicators : A Minority Report In Defense of Causal – Formative Indicators : A Minority Report.

Edwards, J. R. (2011). The Fallacy of Formative Measurement. *Organizational Research Methods*, 14(2), 370–388. <http://doi.org/10.1177/1094428110378369>

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The Problem of Measurement Model Misspecification in Behavioral and Organizational Research and Some Recommended Solutions. *The Journal of Applied Psychology*, 90(4), 710–730. <http://doi.org/10.1037/0021-9010.90.4.710>

Rhemtulla, M., van Bork, R., & Borsboom, D. (2015). Calling Models With Causal Indicators “Measurement Models” Implies More Than They Can Deliver. *Measurement: Interdisciplinary Research and Perspectives*, 13(1), 59–62. <http://doi.org/10.1080/15366367.2015.1016343>

2.3 Reliability

Cortina, J. M. (1993). What Is coefficient Alpha ? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98–104.

McNeish, D. (2017). Thanks Coefficient Alpha, We'll Take It From Here. *Psychological Methods*.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74(1), 145.

Schmitt, N. (1996). Uses and abuses of coefficient Alpha. *Psychological Assessment*, 8(4), 350–353.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107.

2.4 Challenges in the Measurement of Individual Differences

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology*, 8(SEP), 1–16. <http://doi.org/10.3389/fpsyg.2017.01482>

Fröhner, J., Teckentrup, V., Smolka, M., & Kroemer, N. (2018). Addressing the reliability fallacy: Similar group effects may arise from unreliable individual effects. Preprint, 1–29. <http://doi.org/10.1101/215053>

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 1–21. <http://doi.org/10.3758/s13428-017-0935-1>

3. Quantitative Approaches and Models

This section introduces quantitative approaches relevant to measurement in psychology. It starts with factor models that have been used in psychology for many decades, but also introduces other modeling families: item response theory, mixture models, and network models. Like section 2, this section also contains some tutorial papers.

3.1 Factor Analysis

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4(3), 272. doi:10.1037/1082-989X.4.3.272

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 49(2), 78. <http://dx.doi.org/10.1037/cbs0000069>

Meehl, P. E. (1993). Four Queries About Factor Reality. *History and Philosophy of Psychology Bulletin*, 5(2), 4–5.

3.2 Measurement Invariance

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, 44(11).

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures : The issue of partial measurement invariance. *Psychological Bulletin*, 105(3), 456–466.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233–255.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7(4), 361–388.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543.

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4), 486–492.

3.3 Item Response Theory and Differential Item Functioning

* Hambleton, R.K & Russell, W.J. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice* [Module 16: <https://www.ncme.org/ncme/NCME/NCME/Publication/ITEMS.aspx>]

Harris, D. (1989). Comparison of 1-, 2-, and 3-Parameter IRT models. *Educational Measurement: Issues and Practice*. [Module 7: <https://www.ncme.org/ncme/NCME/NCME/Publication/ITEMS.aspx>]

Penfield, R.D. (2014) An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*. 33(1), 36-48. [Module 35: <https://www.ncme.org/ncme/NCME/NCME/Publication/ITEMS.aspx>]

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, 27(4), 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.

3.4 Mixture Models (Latent Class/Profile Analysis)

Finch, W.H. & Bronk, K.C. (2011). Conducting confirmatory latent class analysis using Mplus. *Structural Equation Modeling*, 18, 132-151.

Masyn, K. E. (2013). Latent Class Analysis and Finite Mixture Modeling. In P. Nathan and T. Little (Eds.), *The Oxford Handbook of Quantitative Methods* (pp. 551-611). New York, NY. Oxford University Press.

3.5 Network Models

Epskamp, S., & Fried, E. I. (2018). A Tutorial on Regularized Partial Correlation Networks. *Psychological Methods*. <http://doi.org/10.1037/met0000167>

Foygel, R., & Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. In *Advances in Neural Information Processing Systems 23* (pp. 604–612).

van der Maas, H. L. J., Dolan, C. V, Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review*, 113(4), 842-61. <http://doi.org/10.1037/0033-295X.113.4.842>

4. Analyses of Measurement Practices in Psychology

These papers illuminate what measurement practices and problems are common in different areas of psychological research, and suggest ways to improve them.

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and Reliability Reporting Practices in the Field of Health Education and Behavior. *Health Education & Behavior*, 41(1), 12–18. <http://doi.org/10.1177/1090198113483139>

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychology and Personality Science*, 1–9. <http://doi.org/10.1177/1948550617693063>

Fried, E. I. (2017). The 52 symptoms of major depression. *Journal of Affective Disorders*, 208, 191–197. <http://doi.org/10.1016/j.jad.2016.10.019>

Hubley A.M., Zhu S.M., Sasaki A., Gadermann A.M. (2014) Synthesis of Validation Practices in Two Assessment Journals: Psychological Assessment and the European Journal of Psychological Assessment. In: Zumbo B., Chan E. (eds) *Validity and Validation in Social, Behavioral, and Health Sciences. Social Indicators Research Series*, vol 54.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual review of clinical psychology*, 5, 27-48.

Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J., Bernstein, A., ... Lenze, E. (2016). Unreliability as a Threat to Understanding Psychopathology: The Cautionary Tale of Attentional Bias. *Journal of Abnormal Psychology*, 125(6), 840–851. <http://doi.org/10.1037/abn0000184>

Santor, D. A., Gregus, M., & Welch, A. (2006). Eight Decades of Measurement in Depression. *Measurement*, 4(3), 135–155.

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, 17(2), 267.

5. Examples of Construct Validation Efforts in Substantive Research

Flake, J. K., Barron, K. E., Hulleman, C., McCoach, D. B., & Welsh, M. E. (2015). Measuring cost: The forgotten component of expectancy-value theory. *Contemporary Educational Psychology*, 41, 232–244. <http://doi.org/10.1016/j.cedpsych.2015.03.002>

Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological bulletin*, 136(3), 422.

McCoach, D. B., & Siegle, D. (2003). The school attitude assessment survey-revised: A new instrument to identify academically able students who underachieve. *Educational and Psychological Measurement*, 63(3), 414–429.

Miller, F. G., Johnson, A. H., Yu, H., Chafouleas, S. M., McCoach, D. B., Riley-Tillman, T. C., ... & Welsh, M. E. (2018). Methods matter: A multi-trait multi-method analysis of student behavior. *Journal of School Psychology*, 68, 53–72.

Pastor, D. A., Barron, K. E., Miller, B. J., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*, 32(1), 8–47.

6. Reference Texts and Book Chapters

- Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge University Press. <http://doi.org/https://doi.org/10.1017/CBO9780511490026>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.
- Embretson, S., & Reise, S. (2013). *Item Response Theory*. Psychology Press.
- Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.
- McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain*. New York, NY: Springer.
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.
- Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. Springer.
- Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research* (Vol. 26). New York, NY: Springer.

7. Here be Dragons: Future Sections & Suggestions

We add papers & books to this section 7 when suggestions are made (usually via Twitter and Facebook), and try to find time every few months to discuss these amongst ourselves. We then either integrate or delete the resources. These resources still need to be integrated:

- Self-report vs clinician-report
 - Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist*, 54, 93–105.
- Interviews vs structured interviews
- Retrospective report & recall
- Measurement issues in intensive daily diary & experience sampling studies
- http://www.bwgriffin.com/gsu/courses/edur9131/content/Haynes_Content_Validity_Assessment.pdf
- <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0202873>
- https://www.researchgate.net/publication/230026507_Classical_and_Modern_Methods_of_Psychological_Scale_Construction
- Recommendation Uli: use of multi-method approach to remove error variance from environmental variance in twin studies.
(<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6494.1997.tb00324.x>)
- Recommendation Uli: Theory underlying multi-method measurement of traits
(<http://psycnet.apa.org/record/2010-10285-006>)
- Recommendation Ruben: Maybe this could be added for generalizability of change and related concepts. Shrout, P., & Lane, S. P. (2012). Psychometrics. In T. S. Conner & M. R. Mehl (Eds.), *Handbook of research methods for studying daily life* (pp. 302–320). New York: Guilford Press.
- Simon: Maybe a reference to G theory would be useful addition to the reliability section in providing a broader perspective on what reliability is (i.e., beyond it's relation to internal consistency). [e.g., Cronbach et al., 1963; Shavelson & Webb, 1991]
- Dorota: The Role of Measurement Quality on Practical Guidelines for Assessing Measurement and Structural Invariance
- Kemper, C. J., Trapp, S., Kathmann, N., Samuel, D. B., & Ziegler, M. (2018). Short Versus Long Scales in Clinical Assessment: Exploring the Trade-Off Between Resources Saved and Psychometric Quality Lost Using Two Measures of Obsessive–Compulsive Symptoms. *Assessment*. <http://doi.org/10.1177/1073191118810057>
- Suggested in PsychMap:
<https://www.cambridge.org/core/books/psychology-of-survey-response/46DE3D6F7C1399BCDC78D9441C630372>
- Amazing Millsap intro book to measurement invariance (via Tom Booth)
- Kline "A handbook of test construction"; further: Moreno, R. Martínez, R. Muñoz, J. (2015). GUIDELINES BASED ON VALIDITY CRITERIA FOR THE DEVELOPMENT OF MULTIPLE CHOICE ITEMS. *Psicothema*, 27 (4), 388-394. <https://doi.org/10.7334/psicothema2015.110>