Review Article

BIOMEDICAL Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2023.52.008280

Why Reliability of Trait Scores is Typically Overestimated and What to Do About It

Walter P Vispoel*

Department of Psychological and Quantitative Foundations, University of Iowa, USA

***Corresponding author:** Walter P Vispoel, Department of Psychological and Quantitative Foundations, University of Iowa, USA

ARTICLE INFO

Received: August 25, 2023

Published: August 31, 2023

Citation: Walter P Vispoel. Why Reliability of Trait Scores is Typically Overestimated and What to Do About It. BJSTR. MS.ID.008280.

ABSTRACT

In this article, I illustrate problems of conventional reliability coefficients routinely overestimating overall consistency of scores from trait-based measures along with effective techniques to address these issues. I also direct readers to computer resources that would enable them to apply those techniques.

A Common Problem in Research Studies

Proper use of assessment tools is predicated on assumptions that results are reliable and valid (American Educational Research Association [1]). Reliability represents the extent to which scores from assessment measures provide consistent results and, without adequate consistency, assessments cannot produce valid results. For objectively scored measures, the most common indices of reliability reported in the research literature are based on single occasions of administration. Examples of such indices include alpha (Cronbach [2]), omega (McDonald [3]), and split-half (Spearman [4]) reliability estimates. Although these coefficients are convenient and straightforward to calculate, they have been criticized for misrepresenting and typically overestimating reliability for measures of traits because they do not account for all pertinent sources of measurement error that could affect scores. Omission of important sources of measurement error, in turn, will lead to underestimates of relationships between constructs when those reliability coefficients are used to correct correlation coefficients for measurement error.

Using Generalizability Theory (G-theory) Techniques as a Possible Solution

To overcome such problems, objectively scored measures of traits can be administered on at least two occasions, separated by a short time interval (e.g., one to two weeks). This brief gap between administrations is intended to reduce effects of memory and ensure that levels of traits remain reasonably stable over the interval. Although this procedure would allow for separate calculation of both single-occasion and test-retest coefficients, neither type of coefficient by itself would account for all relevant sources of measurement error. Specifically, single-occasion indices fail to account for inconsistency across occasions, and test-retest coefficients fail to account for inconsistency across items. Generalizability theory (G-theory; Brennan, Cronbach, Shavelson, Vispoel & others, see [5-9]) and related structural equation modeling techniques (see, e.g., Vispoel & colleagues [10-12]) provide mechanisms to account for and separate the key sources of measurement error that affect scores. For objectively scored measures, these types of error are often referred as specific-factor error or method effects, transient error or state effects, and random-response error or "within-occasion noise" effects (see Schmidt, Le, & Ilies [13] for more in-depth discussions of these types of error). To provide readers with examples of the magnitude of such effects within self-report measures, I summarize results from several recent studies conducted by our research group at the University of Iowa involving responses to popular measures of self-concept, personality, and socially desirable responding that were taken by college students twice, a week apart (see Table 1).

Table 1: Average Proportions of Trait and Measurement Error Variance and Reliability Coefficients for Selected Trait-based Measures.

Proportions of Variance and Corresponding Reliability Coefficients							
Domain/Measure	US/ Trait	SFE/ Method	TE/ State	RRE/ Noise	CE (Items Only)	CS (Occasions Only)	CES (Items & Occasions)
Self-concept							
[a]RSES (n = 555)	0.840	0.036	0.076	0.062	0.916(9.05)	0.876(4.29)	0.840
[b]TSBI (n = 206)	0.767	0.068	0.103	0.062	0.870(13.42)	0.835(8.87)	0.767
[c]SDQ-III (n = 427)	0.844	0.044	0.060	0.052	0.904(7.11)	0.888(5.21)	0.844
Personality							
[c]IPIP-BFM-100 (n = 359)	0.846	0.042	0.072	0.040	0.918(8.51)	0.888(4.96)	0.846
[d]BFI (n = 919)	0.750	0.070	0.093	0.087	0.843(12.40)	0.820(9.33)	0.750
[e]BFI-2 Domain Scales (n = 340)	0.802	0.068	0.061	0.068	0.863(7.61)	0.870(8.48)	0.802
[e]BFI-2 Facet Scales (n = 340)	0.664	0.141	0.055	0.140	0.719(8.28)	0.805(21.23)	0.664
Socially desirability responding							
[c]BIDR polytomous (n = 585)	0.696	0.128	0.071	0.104	0.767(10.20)	0.824(18.39)	0.696
[c]BIDR dichotomous (n = 585)	0.674	0.100	0.099	0.127	0.773(14.69)	0.774(14.84)	0.674
[f]PDS polytomous (n = 195)	0.723	0.139	0.044	0.091	0.767(6.09)	0.862(19.23)	0.723
[f]PDS dichotomous (n = 195)	0.634	0.131	0.091	0.144	0.725(14.35)	0.765(20.66)	0.634

Note US = universe-score variance, SFE = specific-factor error/method variance, TE = transient error/state variance, RRE = random-response error/within-occasion noise variance, CE = coefficient of equivalence, CS = coefficient of stability, CES = coefficient of equivalence and stability, RSES = Rosenberg Self-Esteem Scale (Rosenberg [14-15]), TSBI = Texas Social Behavior Inventory (Helmreich & Stapp [16]), SDQ-III = Self-Description Questionnaire-III (Marsh [17]), IPIP-BFM-100 = 100 item version of the International Item Pool Big Five Model questionnaire (Goldberg [18]), BFI = Big Five Inventory (John, Donahue, & Kentle [19]), BFI-2 = recently revised and expanded form of the BFI that includes both personality domain and nested facet scores (Soto & John), BIDR = Balanced Inventory of Desirable Responding (Paulhus [21]), PDS = Paulhus Deception Scales (Paulhus [22]). Values in parentheses in columns 6 and 7 represent the percent to which CEs or CSs exceed CESs.

[a] From Vispoel, Hong, and Lee [10].

[b] From Vispoel, Morris, and Kilinc [9].

[c] From Vispoel, Lee, Chen, and Hong [11].

[d] From Vispoel, Xu, and Kilinc [12].

[e] From Vispoel, Lee, and Hong [23].

[f] From Vispoel, Morris, and Kilinc [24].

Examples of Applications of Generalizability Theory with Self-Report Measures

Self-concept measures represented in Table 1 include the Rosenberg Self-Esteem Scale (RSES; Rosenberg [14-15]), Texas Social Behavior Inventory (Helmreich & Stapp [16]), and Self-Description Questionnaire III (SDQ-III; Marsh [17]); personality measures include the 100-item International Personality Item Pool Big Five Model questionnaire (IPIP-BFM-100; Goldberg [18]), Big Five Inventory (BFI; John, Donahue, & Kentle [19]), and updated Big Five Inventory-2 (BFI-2; Soto & John [20]); and socially desirable responding measures include the Balanced Inventory of Desirable Responding (BIDR; Paulhus [21]) and Paulhus Deception Scales (PDS; Paulhus [22]). For sake of brevity, the results in Table 1 represent averages across all subscales within each instrument. The same information for each individual subscale is provided in the articles cited in the footnotes to the table. The second through fifth columns of Table 1 represent average partitioning of observed score variance for each instrument when administered on a single occasion but with G-theory techniques applied to estimate proportions for each type of measurement error. Note that each source of error accounts for noteworthy proportions of observed score variance for all instruments, ranging on average from 0.036 to 0.141 for specific-factor/method effects, 0.044 to 0.103 for transient error/state effects, and 0.040 to 0.144 for random-response error/within-occasion noise effects. These values would be even higher for some individual scales within the instruments that have multiple subscales. Because subscale scores are combined to form personality domain composite scores in the BFI-2, partitioning

is described separately at both levels. Similarly, because the BIDR and PDS can be scored polytomously or dichotomously to emphasize exaggerated degrees of socially desirable responding, partitioning for both types of scores are presented for those instruments (see Paulhus [21-22]).

When single-occasion reliability estimates for objectively scored measures are reported in research studies, transient error/state effects become part of trait variance (i.e., universe score or person variance in applications of G-theory, true score variance in classical test theory, and communality in factor models). In contrast, when test-retest reliability estimates are reported, specific-factor error/method effects overlap with trait variance. As a result, both single-occasion and test-retest coefficients routinely overestimate the overall reliability of scores. A major benefit of G-theory and related techniques is that multiple sources of measurement error variance can be separated to determine their individual effects on observed scores and generate a variety of reliability (or generalizability) coefficients catered to just item effects, just occasion effects, or both types of effects. Specific-factor error/method effects also can be compared to transient error/state effects to determine the best ways to alter measurement procedures to enhance reliability. Other things being equal, adding items is a better way to improve reliability when specific-factor error/method effects exceed transient/state effects, whereas adding occasions would be more effective when the opposite is true. For example, based on results in Table 1, we would infer that reliability of results from the Rosenberg Self-Esteem Scale (RSES; see [10]) could be better improved by pooling results over additional occasions rather than increasing numbers of items, but the reverse would be true for personality facet scores from the latest version of the Big Five Inventory (BFI-2; see [20]).

The last three columns of Table 1 represent three types of reliability/generalizability coefficients that can be derived from the data collected in the studies cited here. Reliability coefficients specific to items are sometimes called coefficients of equivalence (CEs). In the present examples, these would be analogous to alpha reliability estimates. Note that they can be derived by adding proportions of trait and transient error/state variance, thereby showing explicitly that these sources of variance are confounded within single-occasion reliability estimates. Reliability coefficients based solely on occasions are analogous to test-retest coefficients and are sometimes called coefficients of stability (CSs). These coefficients can be computed by adding proportions of trait and specific-factor error/method variance, highlighting that these two sources of variance are confounded within test-retest coefficients. Finally, the last column in Table 1 represents reliability/generalizability coefficients that take both item and occasion effects into account and are therefore sometimes referred to as coefficients of equivalence and stability (CESs). Note that these coefficients are identical to proportions of trait variance when separating out the three relevant sources of measurement error. Although lower than the previous coefficients, CESs would typically provide more appropriate indexes of reliability for measures of traits. Values in parentheses in columns 6 and 7 of Table 1 indicate the percent to which CEs and CSs exceed CESs on average within each instrument. These values show that overestimation of reliability is a pervasive problem across inventories with CEs exceeding CESs by 6.09% to 14.69% (Mean = 10.16%) and CSs doing so by 4.29% to 21.23% (Mean = 12.32%). Again, these percentages would be even higher for some subscales within multi-scale inventories.

Final Comments

My purpose in writing this short article was to emphasize that reliability of scores from measures of traits is routinely overestimated when researchers report conventional indices of score consistency. G-theory provides one potential way to derive more appropriate estimates of score consistency by accounting for multiple sources of measurement error. The techniques described here also can be readily applied to subjectively scored measures and further fine-tuned to accommodate more complicated relationships between scores and underlying traits, account for effects of item wording, adjust for scale coarseness problems common with binary and ordinal data, build confidence intervals around key parameter estimates, and derive alternative indices of consistency and agreement for criterion-referenced uses of scores. These techniques are covered in depth in the articles cited in Table 1. Most of these articles also have additional online supplements with instructions on how to use readily available software to apply all illustrated techniques. I hope my brief foray into the nature of reliability helps readers in broadening perspectives concerning measurement error effects on scores from trait-based measures and in deriving more appropriate indices of consistency when using and evaluating such measures.

References

1. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). Standards for educational and psychological testing. Washington DC: American Educational Research Association.
2. Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3): 297-334.
3. McDonald RP (1999) Test theory: A unified approach. Mahwah, NJ: Erlbaum.
4. Spearman C (1904) The proof and measurement of association between two things. American Journal of Psychology 15 (1): 72-100.
5. Brennan RL (2001) Generalizability theory. New York, NY: Springer-Verlag.
6. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York, NY: Wiley 178(4067): 1275-1275A.
7. Cronbach LJ, Rajaratnam N, Gleser GC (1963) Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16(2): 137-163.

8. Shavelson RJ, Webb NM (1991) Generalizability theory: A primer. Sage.
9. Vispoel WP, Morris CA, Kilinc M (2018a) Applications of generalizability theory and their relations to classical test theory and structural equation modeling. Psychological Methods 23(1): 1-26.
10. Vispoel WP, Hong H, Lee H (2023) Benefits of doing generalizability theory analyses within structural equation modeling frameworks: Illustrations using the Rosenberg Self-Esteem Scale [Teacher's corner]. Structural Equation Modeling: A Multidisciplinary Journal. Advance Online Publication.
11. Vispoel WP, Lee H, Chen T, Hong H (2023) Using structural equation Modeling techniques to reproduce and extend ANOVA-based generalizability theory analyses for psychological assessments. Psych 5(2): 249-273.
12. Vispoel WP, Xu G, Kilinc M (2021) Expanding G-theory models to incorporate congeneric relationships: Illustrations using the Big Five Inventory. Journal of Personality Assessment 104(4): 429-442.
13. Schmidt FL, Le H, Ilies R (2003) Beyond alpha: An empirical investigation of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. Psychological Methods 8(2): 206-224.
14. Rosenberg M (1965) Society and the adolescent self-image. Princeton University Press.
15. Rosenberg M (1989) Society and the adolescent self-image (revised edition.), Wesleyan University Press.
16. Helmreich R, Stapp J (1974) Short forms of the Texas Social Behavior Inventory (TSBI), an objective measure of self-esteem. Bulletin of the Psychonomic Society 4: 473-475.
17. Marsh HW (1992) Self-Description Questionnaire (SDQ) III: A theoretical and empirical basis for the measurement of multiple dimensions of late adolescent self-concept. An interim test manual and research monograph; University of Western Sydney: Macarthur, Australia.
18. Goldberg LR (1992) The development of markers for the Big-Five factor structure. Psychological Assessment 4(1): 26-42.
19. John OP, Donahue EM, Kentle RL (1991) The Big Five Inventory-versions 4a and 54. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.
20. Soto CJ, John OP (2017) The next Big Five Inventory (BFI-2): Developing and accessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. Journal of Personality and Social Psychology 113(1): 117-143.
21. Paulhus DL (1991) Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.)., Measures of personality and social psychological attitudes. San Diego, CA: Academic Press pp: 17-59.
22. Paulhus DL (1998) Paulhus Deception Scales (PDS): The Balanced Inventory of Desirable Responding-7 (User's manual). Toronto, Ontario, Canada: Multi-Health Systems.
23. Vispoel WP, Lee, H, Hong H (2023) Analyzing multivariate generalizability designs within structural equation modeling frameworks [Teacher's corner]. Structural Equation Modeling: A Multidisciplinary Journal. Advance Online Publication.
24. Vispoel WP, Morris CA, Kilinc M (2018b) Using G-theory to enhance evidence of reliability and validity for common uses of the Paulhus Deception Scales. Assessment 25(1): 69-83.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2023.52.008280

Walter P Vispoel. Biomed J Sci & Tech Res