Short Communication

BIOMEDICAL Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2025.63.009827

Using Generalizability Theory to Estimate the Dependability of Individual Cut Scores within Biomedical and Other Assessments

Walter P Vispoel[1]*, Tingting Chen[1] and Hyeryung Lee[2]

[1]*University of Lowa, USA*

[2]*Oklahoma State University, USA*

***Corresponding author:** Walter P Vispoel, University of Lowa, USA

ARTICLE INFO

Received: August 07, 2025

Published: August 20, 2025

Citation: Walter P Vispoel, Tingting Chen and Hyeryung Lee. Using Generalizability Theory to Estimate the Dependability of Individual Cut Scores within Biomedical and Other Assessments. Biomed J Sci & Tech Res 63(1)-2025. BJSTR. MS.ID.009827.

ABSTRACT

In this brief article, we describe how generalizability theory can be used to derive indices representing the dependability of cut scores used in decision making based on data from one or two occasions. We include an empirical example using the Neuroticism subscale from the Big Five Inventory to illustrate use of these coefficients and how the dependability of cut scores can exceed the overall dependability of scores when considered on whole. We also direct readers to resources with computer code and further information about conducting the illustrated analyses.

Introduction

Proper interpretation of results from measurement procedures used in biomedicine and other fields depends heavily on the accuracy of scores obtained from those procedures. Common indices of reliability used in practice include alpha, omega, and split-half coefficients based on single occasions and test-retest coefficients over occasions. In general, these coefficients are reported for norm-referencing purposes such as rank ordering in which scores are compared across individuals. However, in many instances, the purpose of an assessment is based on criterion-referencing in which absolute levels of scores takes precedence. Such decisions are typically based on cut scores that reflect different defined levels of behavior, proficiency, medical conditions, and other attributes (see, e. g., Nitko [1]).

Using Generalizability Theory to Derive Indices of Score Accuracy

Generalization theory (G-theory) [2-5] provides a framework in which indices of score accuracy can be derived for either norm- or criterion-referencing purposes that are, respectively, referred to as generalizability and dependability coefficients. Like conventional reliability estimates, G-theory based generalizability and dependability coefficients can range from 0 to 1, with higher values representing greater score accuracy. Two types of dependability coefficients can be reported that both reflect the combined accuracy of scores in relation to their relative and absolute differences [6,7]. The first index represents the overall dependability of scores across the assessment continuum, whereas the second is catered to specific score values. This latter type of dependability coefficient is especially valuable when cut scores are used for decision making at targeted levels of attributes. Such a coefficient in practice can be derived for all possible score values but is often catered to a subset of scores that are used for selection, classification, flagging, and other attribute-specific interpretations.

Values of these coefficients reflect consistency in overall score locations in relation to the cut scores over random replications of the measurement procedure and research design. Unlike conventional reliability estimates (e.g., alpha, omega, split-halves, test-retest), dependability indices within a generalizability theory framework also can be adjusted to take multiple sources of measurement error into account. To illustrate the use of these coefficients, we provide an example using live assessment data obtained from 1,165 college students who completed the Neuroticism subscale from the Big Five Inventory (BFI) [8] on two occasions, a week apart. This scale consists of eight items answered along a 5-point Likert-style response metric (1 =Disagree Strongly, 2 = Disagree a Little, 3. = Neutral, 4 = Agree a Little, 5 = Agree Strongly) that is intended to measure overall levels of anxiety, moodiness, and emotional instability with possible score values ranging from 8 to 40. We report indices here for the first occasion that account for measurement error related to item differences alone, and indices based on both occasions that account for occasion as well as item differences. These same techniques can be readily applied to physiological and virtually any other type of quantitative assessment procedure. The values for generalizability and global dependability for the single-occasion data, respectively, equal 0.852 and 0.807. The global dependability is lower than the generalizability coefficient because it accounts for absolute in addition to relative differences in scores.

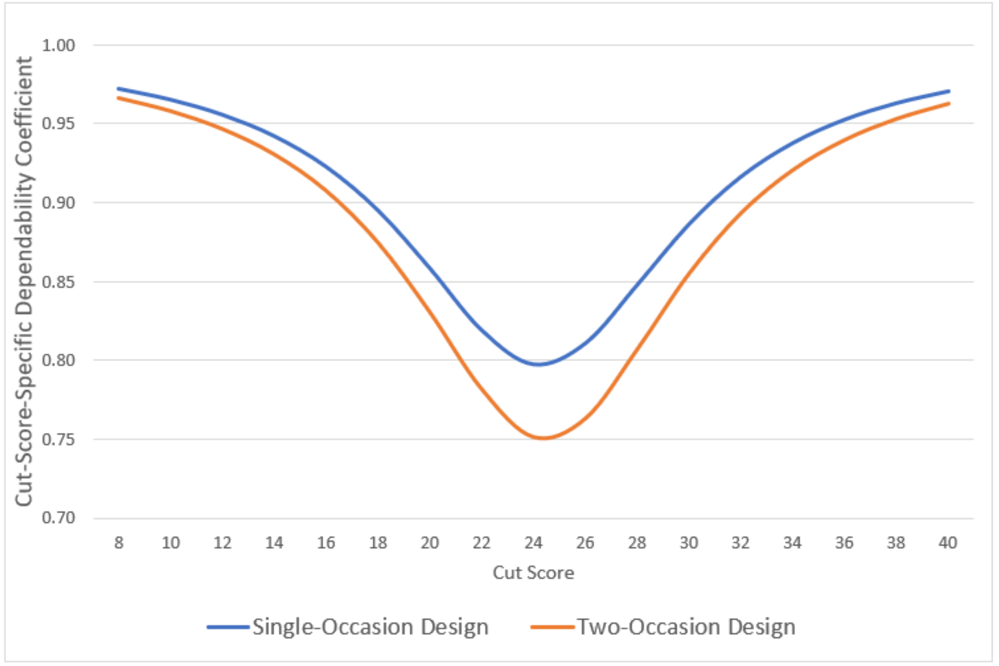

Figure 1: Cut-Score-Specific Dependability Coefficients for the BFI Neuroticism Subscale.

Figure 1 depicts cut-score-specific dependability coefficients for all possible scores from the BFI Neuroticism scale. The cut-score-specific dependability coefficient is lowest at the mean of the score distribution and gradually increases as cut scores deviate away from the mean. When complete population data are available, the minimum cut-score-specific dependability coefficient value would coincide with the global dependability coefficient. If a score of 37 is targeted for possible clinical invention using the single-occasion data, its dependability coefficient of 0.958 noticeably exceeds the global coefficient reported earlier (0.807). This illustrates that the value for a cut-score-specific coefficient can be much greater than the global coefficient when scores are considered on whole. Values of generalizability and global dependability for the two-occasion data, respectively, equal 0.804 and 0.763. compared to 0.852 and 0.807 for the single-occasion data. The reduced values for the two-occasion data occur because they reflect measurement error due to occasion as well as item differences. This relationship also holds for the cut-score-specific dependability coefficients shown in Figure 1. However, the overall trend of increasing values beyond the mean is present with both the single- and two-occasion designs, with the differences between designs diminishing in a similar fashion. For example, the difference between dependability coefficients for the one- and two-occasion data equals 0.046 with a cut score of 24 (approximately at the scale mean) versus 0.011 with a cut score of 37 (roughly two standard deviations above the scale mean).

Conclusion

Our brief illustration here is intended to highlight the importance of catering indices of score accuracy to the purpose of an assessment and the value of G-theory in providing such indices for norm- and criterion-referencing interpretations of scores while taking just item or both occasion and item differences together into account. More detailed information with additional examples and computer code for implementing these procedures can be found in Vispoel, et al. [9-11]. Illustrations of using G-theory with physiological measures appear in a recent article by Clayson [12].

References

1. Nitko A J (1980) Distinguishing the many varieties of criterion-referenced tests. Review of Educational Research 50(3): 461-485.
2. Brennan RL (2001) Generalizability theory. Springer-Verlag.
3. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N (1972) The dependability of behavioral measurements: Theory of generalizability for scores and profiles. Wiley.
4. Shavelson RJ, Webb NM (1991) Generalizability theory: A primer. Sage Publications Inc.
5. 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.
6. Brennan RL, Kane MT (1977) An index of dependability for mastery tests. Journal of Educational Measurement 14: 277-289.
7. Kane MT, Brennan RL (1980) Agreement coefficients as indices of dependability for domain-referenced Tests. Applied Psychological Measurement 4(1): 105-126.
8. John OP, Donahue EM, Kentle RL (1991) The Big Five Inventory - (BFI) [Database record]. APA PsycTests.
9. Vispoel WP, Hong H, Lee H, Jorgensen TD (2023) Analyzing complete generalizability theory designs using structural equation models. Applied Measurement in Education 36(4): 372-393.
10. Vispoel WP, Lee H, Chen T, Hong H (2023) Using structural equation modeling to reproduce and extend ANOVA-based generalizability theory analyses for psychological assessments. Psych 5(2): 249-272.
11. Vispoel WP, Morris CA, Kilinc M (2018b) Practical applications of generalizability theory for designing, evaluating, and improving psychological assessments. Journal of Personality Assessment 100(1): 53-67.
12. Clayson PE (2024) The psychometric upgrade psychophysiology needs. Psychophysiology replace with 61(3): e14522.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2025.63.009827

Walter P Vispoel. Biomed J Sci & Tech Res