SMGroup

SM Journal of Biometrics & Biostatistics

Research Article

# Using Generalizability Theory (G-Theory) to Examine the Reliability of Body Composition Measurement

**Peter D Hart**[1,2,3]*

[1]*Health Promotion Program, Montana State University-Northern, USA*
[2]*Kinesmetrics Lab, Montana State University-Northern, USA*
[3]*Health Demographics, USA*

**Article Information**

Received date: Nov 01, 2017
Accepted date: Nov 20, 2017
Published date: Nov 22, 2017

***Corresponding author**

Peter D Hart, Health Promotion Program, Montana State University-Northern, USA, Tel: 406.265.4129; Fax: 406.265.4129; Email: peter.hart@msun.edu



**Keywords** Generalizability theory; Body composition; Measurement; Biometrics

## Abstract

**Purpose**: Adequate reliability of Body Composition (BC) assessment is a requirement before such measures can be considered valid. Many studies to date have only examined a single source of measurement error such as that from trials (test-retest). Generalizability Theory (G-theory) is a statistical technique that allows for the examination of different sources of measurement error simultaneously in a single analysis. Therefore, the aim of this study was to examine the different sources of error seen in the assessment of BC. A secondary purpose was to determine the appropriate number of facet conditions required to gain a reliable BC measure.

**Methods**: This measurement study included 38 participants who had been assessed on two different occasions (in the same week) and on each of four different BC field methods: Percent Body Fat (PBF) by Skinfold Technique (SF), Waist Circumference (WC), Body Mass Index (BMI) and PBF by Hand-Held Bioelectrical Impedance (HH). Two different G-theory designs were used in this research. First, a two-facet crossed p×t×m design was analyzed treating all facets as random. Then, the same design was performed treating BC method as a fixed facet. In both designs, a Generalizability Study (G-study) and Decision Study (D-study) were conducted. Three different software packages were used to ensure consistent and valid results (GENOVA, SPSS macro, and SAS GLM).

**Results**: The completely random design showed the largest variance component for persons (p) (57.8%). Variance components for both trials (t) and BC method (m) were negligible. However, the interaction between person and method (p×m) was substantial (38.6%). D-study results indicated reliable BC scores for measurement designs administered once using three different methods (G=.803). The mixed design, averaging over BC method, showed majority of variance due to persons (98.5%) and each of the four BC methods showed reliable scores with a single trial (G's>.945).

**Conclusion:** Results from this G-theory research indicate that the equivalence reliability of commonly administered BC assessments may be inadequate. Although different BC assessments individually are reliable, for dependable BC trait generalization to the universe, a minimum of three different methods administered once may be required.

## Introduction

Body composition (BC) usually refers to either an absolute amount of fat and fat-free mass within the body or an amount relative to total body weight [1]. BC receives a lot of attention in the health sciences because of its relationship with premature disease and mortality as well as its increasing prevalence [2]. Moreover, BC standards are promoted in the health sciences because it's one of the five health-related components of fitness, along with cardiorespiratory fitness, muscular strength, muscular endurance, and flexibility [3]. With an increased interest in BC comes increased attention placed on its various means of assessment. Although many laboratory methods exist to assess an individual's BC (e.g., hydrostatic weighing, dual-energy x-ray absorptiometry and air displacement plethysmograph), it is more commonly assessed using field-based techniques [4].

While field-based assessments allow for a lower-cost and more time efficient means of BC assessment, they also present more opportunity to measure a trait with error [5]. This type of error introduced into the assessment process reduces the reliability of the measurement. Furthermore, in order for a measurement process to be considered valid, it must first be considered reliable. In other words, adequate reliability of BC assessment is a requirement before BC measures can be considered valid [6]. Many studies have shown acceptable reliability among BC assessments; however, these studies have only examined a single source of measurement error such as that from trials (test-retest) [7].

Generalizability theory (G-theory) is a statistical technique that allows for the identification and estimation of different sources of measurement error [8]. These different sources of error (e.g., item, occasion, rater, test form) are called facets of a measurement. In a simple Generalizability Study (G-study), say person-by-test form, we design a measurement process that will allow us to

isolate and estimate the measurement error attributed by test form (facet). In this G-study example, person is considered the object of measurement and therefore not a facet. A decision study (D-study) is often conducted using variance information from the G-study and can result in a new measurement procedure that minimizes measurement error (e.g., 3 test forms required to achieve desired reliability).

G-theory has been used to examine the reliability of several different measurement procedures. One study used D-study results to determine the number of accelerometer wear days needed to obtain reliable estimates of physical activity and sedentary behavior in 9 to 11 year children [9]. Another study used G-theory methods to identify the measurement error associated with the number of different scenarios applied to taping skills using an athletic training assessment instrument [10]. G-theory and D-study results have also been applied to the self-monitoring of blood pressure to determine the number of self-taken readings needed to get reliable estimates of both systolic and diastolic pressure [11]. Finally, G-theory has been used extensively in validity and reliability studies of various patient-reported outcome tools [12-14]. To date, however, G-theory has not been used to examine the assessment of BC across several different methods.

Given this background regarding BC and G-theory, the aim of this study was to examine the different sources of error seen in the assessment of BC. A secondary purpose was to estimate the appropriate number of facet conditions (e.g., number of trial, number of methods) needed to gain reliable measures.

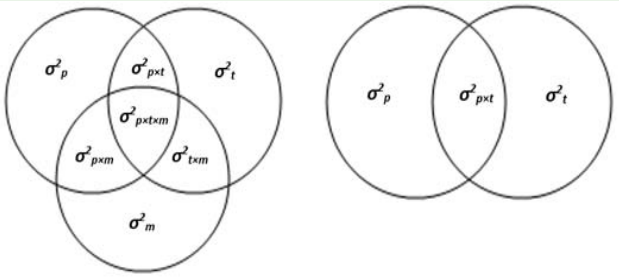
## Methods

### Participants

Data for this research came from a cross-sectional measurement study conducted at a rural public university. Participants were recruited by both study flyer and word-of-mouth. A total of N=38 college students who had their BC assessed by all four methods were included in the study. All study components were reviewed and approved by the university's Institutional Review Board (IRB).

### Research Design

This study used a fully crossed repeated measures design with participants assessed on two different occasions (in the same week) on each of four different BC field methods. Figure 1 displays the two different G-theory designs.




Completely crossed two facet design treating BC method as random

Completely crossed single facet design treating each BC method separately

**Figure 1:** Sources of variance for both G-theory designs.
**Note:** Person (*p*), Trial (*t*) and Method (*m*).

**Table 1:** Descriptive statistics and correlations for BC measures (N=38).

| | Trial 1 | | Trial 2 | | Paired t |
|---|---|---|---|---|---|
| **BC measure** | *M* | *SD* | *M* | *SD* | *p* |
| **SF** | 18.3 | 7.68 | 18.4 | 7.73 | 0.766 |
| **WC** | 81.9 | 10.05 | 81.5 | 9.6 | 0.453 |
| **BMI** | 25.4 | 3.68 | 25.5 | 3.59 | 0.715 |
| **HH** | 25 | 9.07 | 25.2 | 8.7 | 0.558 |

**Note:** M is mean. SD is standard deviation. Paired t is paired t statistic.

### Body composition measures

Four different BC measures were used in this study: Percent Body Fat (PBF) by Skinfold Technique (SF), Waist Circumference (WC), Body Mass Index (BMI) and PBF by Hand-Held Bioelectrical Impedance (HH). BC measures were assessed in a laboratory by trained research assistants. PBF (%) by SF was measured using the Siri equation and the sum of chest, abdomen and thigh skinfolds for males and triceps, suprailiac and thigh skinfolds for females [9]. WC (cm) was measured the same for males and females using an elastic tape at the most narrow point between the xyphoid process and umbilicus [9]. PBF (%) by HH was measured using the Omron BF306 handheld bioelectrical impedance device as described by the manufacturer [10]. Finally, BMI (kg/m$^2$) was measured the same for males and females by measuring height (cm) using a wall mounted stadiometer and weight (kg) using an electronic floor scale [9].

### Statistical analysis

Two different G-theory designs were used in this research [11]. First, a two-facet crossed p×t×m design was run treating all facets as random. Then, the same design was performed treating BC method as a fixed facet. In both designs, a Generalizability Study (G-study) and Decision Study (D-study) were conducted. Each BC variable was T-score transformed by sex prior to analysis. Three different software packages were used to ensure consistent results: GENOVA [12], SPSS macro [13] and SAS GLM [14].

## Results

Table 1 displays descriptive statistics for each of the four BC methods across the two trials. No significant differences were noted across trial, which supports the stability of the BC methods. Table 2 displays results for the two facet completely random design.

**Table 2:** G-Study and variance components for two facet crossed *p×t×m* fully random design.

| Source | MS | Levels | σ² Component | % |
|---|---|---|---|---|
| **Person (*p*)** | 533.2 | 8 | 56.674 | 57.8 |
| **Trial (*t*)** | 10.2 | 152 | 0.059 | 0.1 |
| **Method (*m*)** | 0.6 | 76 | 0 | 0 |
| ***pxt*** | 4.1 | 4 | 0.231 | 0.2 |
| ***pxm*** | 78.9 | 2 | 37.849 | 38.6 |
| ***txm*** | 0.3 | 38 | 0 | 0 |
| ***pxtxm*** | 3.2 | 1 | 3.194 | 3.3 |

**Note:** Levels is the number of levels the scores were summed across in that factor. m and *t×m* had negative σ² components and so set to zero after all calculations. % is the percentage of total variance.

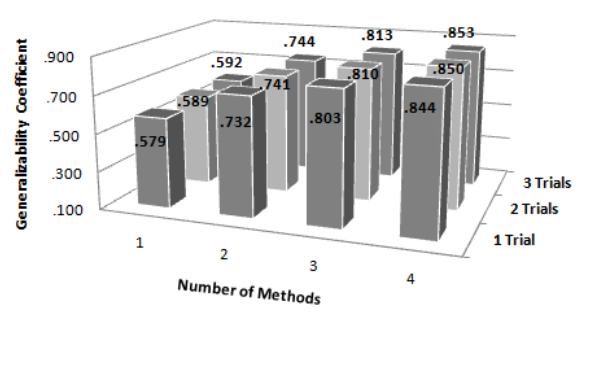



**Figure 2:** G-Study coefficients for two facet crossed *p×t×m* design.

**Table 3:** G-Study and variance components for two facet crossed *p×t×m* mixed design treating method (*m*) as fixed.

| Source | MS | Levels | $\sigma^2$ Component[a] | % |
|---|---|---|---|---|
| **Person (p)** | 533.2 | 8 | 66.136 | 98.4 |
| **Trial (t)** | 10.2 | 152 | 0.059 | 0.1 |
| **pxt** | 4.1 | 4 | 1.03 | 1.5 |

**Note:** Levels is the number of levels the scores were summed across in that factor. % is the percentage of total variance[a], $\sigma^2$ component is average over the fixed facet (m).

The largest variance component was seen for Person (p) (57.8%). Variance components for Both Trial (t) and BC Method (m) were negligible. However, the interaction between person and method (p×m) was substantial (38.6%). Figure 2 displays results of the D-study. Many different scenarios presented resulting in adequate reliability. However, the more realistic scenario indicated reliable BC scores for measurement designs administered once using three different methods (G=.803).

Table 3 displays results for the mixed design, averaging over BC method. Results showed majority of variance due to person (98.4%). The small residual component (1.5 %) suggests that few trials are needed to generalize scores averaged over BC method (m). Finally, Table 4 shows results indicating that each of the four BC methods have reliable scores with a single trial (G's≥.945). SEM values were clinically small for each individual BC method.

## Discussion

The aim of this study was to employ G-theory in order to examine the different sources of error seen in the assessment of BC.

**Table 4:** Summary of psychometric data for analyses treating each fixed facet (BC method) separately.

| Method | $\sigma^2$ Component | % | G | SEM |
|---|---|---|---|---|
| **SF** | 93.186 | 96.6 | 0.966 | 1.8 |
| **WC** | 92.197 | 94.5 | 0.945 | 2.3 |
| **BMI** | 98.38 | 98.9 | 0.989 | 1 |
| **HH** | 94.329 | 95.9 | 0.959 | 2 |

**Note:** $\sigma^2$ component is universe score variance. % is the percentage of total variance. G is generalizability coefficient for 1 trial. SEM: Standard Error of Measurement.

The purpose in doing so was to further examine the accuracy of generalizing an individual's observed BC test score to an average BC score given under all possible conditions (facets). That is, we might ask, would an individual's BC score assessed at one occasion using one single method be matched if assessed again on a different occasion by a different method? The status quo in the health sciences is to take results from observed scores and generalize them across those facets, without regard to their accuracy. Results of this study clearly indicate that the answer to the above question is "no".

Specifically, results from the two-facet G-study showed a substantial amount of person-by-method (p×m) error variance. This can be interpreted to mean that large differences exist in the relative standing of individuals across the different BC methods. Said in a simpler way, much error was seen when college students were separately ranked in each BC assessment method. This finding would question the equivalence reliability of the different BC methods [5]. Notwithstanding, the notion that different BC methods may rank individuals differently is not entirely unheard of. Individuals with higher fitness levels tend to have more muscle mass as compared to their lower fitness counterparts [15]. Consequently, individuals with more muscle mass are likely to rank low in terms of PBF, however, rank high in terms of BMI. This also seems appropriate given that college-aged individuals possess higher fitness levels than older populations [16].

Using the quantified G-study variance components, results from the two-facet D-study showed that at least three BC assessment methods would be required to yield reliable BC measures (≥.80) in the future. The D-study results also indicated that only a single trial, with three BC methods, would be required. These findings are consistent with the G-study results, since the person-by-method (p×m) error variance was the largest source of error variance [14].

The G-study using a mixed design (i.e., treating trial as a random facet and method as a fixed facet) was conducted for two main reasons. Firstly, to serve as an alternative strategy if one wanted to consider that the four BC methods in this study were not randomly drawn from a population of several different BC methods. Said differently, treating BC method as a fixed facet implies we are unwilling to replace these four BC methods with four other randomly selected methods. Regardless of viewpoint on this matter, the mixed design was included simply as an alternative view. Secondly, given the fully random design results, it made sense to further examine the measurement properties of the BC scores averaging over BC method.

The mixed design G-study results were very clear, in that, almost all variance in BC scores were due to person. This means that our sample of individuals differed systematically (without error) in their BC scores. At the same time, the very small person-by-trial (p×t) (and residual) variance component indicates that individuals had similar relative standings across trials. Consequently, very few trials would be required to generate acceptable reliability in future scenarios.

The results of this current study should be considered along with limitations. For example, participants in this study were registered college students attending a rural public university. Since the measurement properties of a test are situation specific, then these results should be considered only for this specific population [5]. One other limitation in this study was the lack of other considerable sources

of measurement error variance (facets).One additional and plausible facet to consider in the assessment of BC would be the clinician (i.e., research assistant) involved in the measuring process. The addition of this facet could help identify if BC ranking of individuals differed also by clinician, which is a reasonable suggestion. The addition of this facet could also help understand and quantify additional nuances such as if clinicians tended to measure BC differently across BC method. Acknowledging this limitation, it is also worth mentioning that the measurement study for a three-facet (adding clinician) fully crossed design would be very complicated, both for the clinician and the participant. A future study, however, may want to consider such a design; either fully crossed or nested [6].

## Conclusion

Results from this G-theory research indicate that the equivalence reliability of commonly administered BC assessments may be inadequate. Although different BC assessments individually are reliable, for dependable BC trait generalization to the universe, a minimum of three different methods administered once may be required.

### References

1. Kraemer WJ, Fleck SJ, Deschenes MR. Exercise physiology: integrating theory and application. Lippincott Williams & Wilkins. 2011.
2. Centers for Disease Control and Prevention (CDC. Vital signs: state-specific obesity prevalence among adults-United States, 2009. MMWR. Morbidity and mortality weekly report. 2010; 59: 951.
3. American College of Sports Medicine, editor. ACSM's health-related physical fitness assessment manual. Lippincott Williams & Wilkins. 2013.
4. McArdle WD, Katch FI, Katch VL. Exercise physiology: nutrition, energy and human performance. Lippincott Williams & Wilkins. 2010.
5. Morrow JR, Mood D, Disch J, Kang M. Measurement and Evaluation in Human Performance, 5E. Human Kinetics. 2015.
6. Strube MJ. Reliability and generalizability theory. In: Grimm LG, Yarnold PR. Reading and understanding multivariate statistics. American psychological association. 2000.
7. Aandstad A, Holtberget K, Hageberg R, Holme I, Anderssen SA. Validity and reliability of bioelectrical impedance analysis and skinfold thickness in predicting body fat in military personnel. Military medicine. 2014; 179: 208-217.
8. Webb NM, Shavelson RJ. Generalizability theory: Overview. Wiley Stats Ref: Statistics Reference Online. 2005.
9. Barreira TV, Schuna JM, Tudor-Locke C, Chaput JP, Church TS, Fogelholm M. Reliability of accelerometer-determined physical activity and sedentary behavior in school-aged children: a 12-country study. International journal of obesity supplements. 2015; 5: S29-S35.
10. Lafave MR, Butterwick DJ. A generalizability theory study of athletic taping using the technical skill assessment instrument. Journal of athletic training. 2014; 49: 368-372.
11. García-Vera MP, Sanz J. How many self-measured blood pressure readings are needed to estimate hypertensive patients' "true" blood pressure? Journal of behavioral medicine. 1999; 22: 93-113.
12. Ramsenthaler C, Gao W, Siegert RJ, Schey SA, Edmonds PM, Higginson IJ. Longitudinal validity and reliability of the Myeloma Patient Outcome Scale (MyPOS) was established using traditional, generalizability and Rasch psychometric methods. Quality of Life Research. 2017; 26: 2931-2947.
13. Meng Q, Yang Z, Wu Y, Xiao Y, Gu X, Zhang M, et al. Reliability analysis of the Chinese version of the Functional Assessment of Cancer Therapy-Leukemia (FACT-Leu) scale based on multivariate generalizability theory. Health and quality of life outcomes. 2017; 15: 93.
14. Bravo G, Sene M, Arcand M. Reliability of health-related quality-of-life assessments made by older adults and significant others for health states of increasing cognitive impairment. Health and quality of life outcomes. 2017; 15: 4.
15. American College of Sports Medicine. ACSM's guidelines for exercise testing and prescription. Lippincott Williams & Wilkins. 2013.
16. Omron Fat Loss Monitor. Model HBF-306. Omron Healthcare Co., Ltd. 2012.
17. Safrit JM, Wood TM. Measurement in physical education and exercise science. Times Mirror/Mosby College Publishing, St. Louis. 1990.
18. Brennan RL. Generalizability theory. Educational Measurement: Issues and Practice. 1992; 11: 27-34.
19. Mushquash C, O'Connor BP. SPSS and SAS programs for generalizability theory analyses. Behavior research methods. 2006; 38: 542-547.
20. Shavelson RJ, Webb NM. Generalizability Theory: A Primer: A Primer. Sage Publications. 1991.
21. Raven P, Wasserman D, Squires W, Murray T. Exercise Physiology. Nelson Education. 2012.
22. Milanović Z, Pantelić S, Trajković N, Sporiš G, Kostić R, James N. Age-related decrease in physical activity and functional fitness among elderly men and women. Clinical interventions in aging. 2013; 8: 549-556.