Id: 1203729

Participatory Educational Research (PER)
Vol.10(1), pp. 178-189, January 2023
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.23.10.10.1

Examining Cross-Cultural Applicability via Generalizability Theory

Sümeyra SOYSAL*
Department of Educational Sciences, Necmettin Erbakan University, Konya, Turkey
ORCID: 0000-0002-7304-1722

Article history
Received:
24.07.2022

Received in revised form:
30.10.2022

Accepted:
21.11.2022

Key words:
Cross-cultural applicability; generalizability; multi-faceted analysis; G-theory

Applying a measurement instrument developed in a specific country to other countries raise a critical and important question of interest in especially cross-cultural studies. Confirmatory factor analysis (CFA) is the most preferred and used method to examine the cross-cultural applicability of measurement tools. Although CFA is a sophisticated technique to investigate various equivalence types (structural, metric, scalar and alike.), it has some limitations. In light of the classical test theory, when a measurement tool is not invariant between countries, what factors contribute to the error variance become unclear. Also, CFA reveals little as to how dimensionality of the relevant measurement tool affects measurement invariance. Hence, a fundamental focus of this study is to examine the measurement comparability or cross-cultural applicability for different countries on an international assessment using generalizability theory (G-theory) in educational science studies. With multi-faceted design, the contribution of dimensionality to error variance is examined, as well. For illustration purposes, eight scales from PISA 2012 student questionnaire dataset related to attitudes towards mathematics are used. The study is based on data from Türkiye, Finland and USA. The unbalanced multi-faceted designs are performed using G String IV. In conclusion, almost all results supported all research expectations. From the estimations of the G-theory, it can be rightly deduced cross-nationally applicability of the attitudes towards mathematics scales from these research findings.

Introduction

To expand theories and their related concepts or constructs developed in a specific culture to other cultures, a considerable necessity is to evaluate the degree of a given measurement tool's cross-national applicability or equivalence (Van de Vijver et al., 2021, p.47). According to Horn and McArdle (1992), the term "applicability" or "equivalence" means to “whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute” (p.117). Stated in a different way, can the measurement tools, whose consistency and accuracy have already been established, be generalized or applicable to other countries? Applicability implies that the operational definition and notional sense of a specific construct are the identical between

[*] Correspondency: sumeyrasoysal@hotmail.com

nations or countries (Van de Vijver & Leung, 1997, pp.259-260). Also, Matsuma and Van de Vijver (2011, p.19) indicate that equivalence refers to the level of comparability of measurement outcomes. As a result, meaningful cross-national comparisons can only be made if the data or measures from different cultures are comparable (Van de Vijver & Leung, 1997, p.261).

The conventional method preferred by researchers for considering cross-cultural generalizability of measures is confirmatory factor analysis (CFA) whose framework suggested by Steenkamp and Baumgartner (1998). According to Van de Vijver and Leung (2010, p.33), CFA has some limitations as well as its strengths, such as the flexibility to test various equivalence types (structural, metric, scalar etc.) and the potential of examining hierarchically nested models. One of its weaknesses is that the use of CFA in testing equivalence (or invariance in the terminology of structural equation modelling) is encumbered by troubles with fit statistics, especially chi-square, because of sensitivity to large sample sizes. If a poor fit is revealed, it is generally unclear if the problem is because of misspecifications of the latent construct or due to secondary cross-national differentiations that are psychologically insignificant. The reasonably large sample size requirement becomes even more prominent because of requiring the estimation of the CFA model for each group separately. Another its weakness is that it partitions variation into just two resources: true score variance and error score variance because CFA is a method based the classical test theory (CTT) (Shavelson & Webb, 1991). However, this assumption of CTT brings with two unresolved issues about cross-cultural measurement. First, if a measurement tool is not invariant across countries, what factors contribute to the error variance is unclear. Second, CFA reveals little as to how the dimensionality of the measurement tool affects measurement invariance (Durvasula et al. 2006).

Sharma and Weathers (2003) and Durvasula et al. (1993, 2006) have proposed the use of generalizability theory (G-theory) as a different way to examine the applicability of measures between countries. Because any sample sizes are convenient for G-studies and G-theory can divide observed score variance into a lot of different sources: item, person, rater, task, country, and the like and any interactions among those sources. For example, when a measurement tool finds not to be applicable across countries, it is a challenge to identify what causes that defect. Then, is the defect because of differentiations in countries or is a significant interaction between countries and the other sources? CFA is not able to provide such detailed diagnosis, but G-theory is. The number of studies examining and investigating the utility of G-theory in cross-cultural generalizability or applicability is limited, and almost all are related to the field of business, marketing, or management (Durvasula & Lysonski, 2016; Durvasula et al., 2006; Eisend, 2009; Malhotra &Sharma, 2008; Sharma &Weathers, 2003).

Concern with comparing student performance across countries has increased attention recently to international large-scale assessments such as the Progress in International Reading Literacy Study (PIRLS) and the Programme for International Student Assessment (PISA) and The information from international assessments, and the comparisons between educational systems that it invites, plays an increasingly important role in decision-making of policy-makers, and reforms at global, regional and national levels (Johansson, 2016). In analyzing the discourses and data surrounding international large-scale assessments, one of the major challenges is the cross-cultural accuracy and applicability of all measurement tools. Hence, the main purpose of this study is to represent the usage of G-theory in this context, cross-cultural applicability of measures, for the profit of educational science practitioners or researchers. Therefore, this study will seek what causes the variation of measures and examine cross-cultural applicability in the context multidimensionality via generalizability theory. For illustration purposes, the data from attitudes towards mathematics scales in the PISA 2012 Student Questionnaire was used. In addition to the methodological contribution, the other contribution of this work is to research whether attitudes towards mathematics differ over students, cultures, items, and dimensions, and the degree to which they differ, i.e., how generally applicable the cross-cultural admissibility of attitudes towards mathematics is when considering those further variables. Before methodology, a short discussion of the G-theory (for detail, e.g., Brennan, 2001a; Shavelson &Webb, 1991) and the research expectations that were built up for cross-cultural applicability and validity is provided in the following.

Generalizability Theory

Cronbach et al. (1963) have explained the concept of G-theory as follow: "an investigator asks about the precision or reliability of a measure because he/she wishes to generalize from the observations in hand to some class of observations to which it belongs... For example, to ask about the reliability of an essay-examination score is to ask how representative this is of grades that might be given to the same paper by other raters, or of grades on other papers by the same subject." (p. 144).

In psychometry, any observed test score variance (σ_O^2) could be imagined as the composite of the theoretical components- a true score variance (σ_T^2) and a random error variance (σ_E^2) - in CTT. The theory is expressed as follows (Crocker & Algina, 1986, p.114):

$$\sigma_O^2 = \sigma_T^2 + \sigma_E^2 \qquad (1)$$

As error variance can be caused by different conditions, values or coefficients of reliability can differ accordingly. For example, coefficient of stability considers only time differentiation, and coefficient of internal consistency consider differentiation only because of item sampling as random error (Crocker & Algina, 1986, p.117). Because it is assumed that the error sources are independent of each other, different reliability coefficients by operating different descriptions of true and error scores are defined under CTT framework. However, this assumption prohibits the estimations of the potential interactions between different sources. G-theory removes restrictions on classical theory by ensuring procedures that permit a researcher to understand and solve multiple sources of error that promote to σ_E^2. This is answered in part thanks to the administration of some ANOVA methods. According to Brennan (2001a, p.2), in a sense, CTT and ANOVA can be viewed as the parents of G-theory. But G-theory is not merely the conjunction of classical theory and ANOVA and has an original conceptual framework.

Factors in analysis of variance (ANOVA) is called facets in the G-theory. (Sharma & Weathers, 2003). Facets are separated into two kinds. One of two is the facet of generalization which contributes to undesirable variance in observed responses, hereby the measurement tools (e.g., scales, questionaries) must be modelled as to reduce variance arising from these factors, e.g. to diminish variance that originates from differentiations in the comprehension of the items. Otherwise, measurement tool reliability would decrease. As for measurement tool dimensions, one would expect that dimensions discriminately measure diverse parts of a construct. However, the dimensions may not represent the common or same latent construct when the dimension contributes a considerable rate of the total variance (Durvasula & Lysonski, 2016). The other one of two is the facet of differentiation (also called as object of measurement) which contributes to desirable variance. For instance, in any cross-cultural research, it is quite usual to compare and examine how people from different countries answer to a measurement tool. In this situation, persons and countries might become as objects of measurement. Accordingly, measurement tools must be modelled to strengthen variance from facets of differentiation (Durvasula et al., 2006). These facets (students and countries) represent measurement objects that I desire to be compared in the present cross-cultural study.

In the current study, two multi-facet G-theory designs were employed: the S:C×I:D and the S:C×I designs. These designs involve students (S), countries (C), items (I) and dimensions (D). The two factors, countries (C) and dimensions (D), are completely crossed because all students, irrespective of their country, response the same dimensions. Whereas the student is nested within the country (S:C) because of the students from different countries. Correspondingly the item is nested within the dimension (I:D) as items of various dimensions are different. Accordingly, these designs are considered mixed because they have a combination of crossed and nested facets. The variance components and influences of them on generalizability are summarized in Table 1 (for further details, Sharma & Weathers, 2003).

Research expectation

To support cross-cultural generalizability or applicability in multi-facet analysis (S:C×I:D) design, the following research expectations developed by some researchers (Durvasula & Lysonski, 2016; Durvasula et al., 2006; Malhotra & Sharma, 2008) also summarized in Table 1, need to be confirmed:

(1) *Variance accounted for by between country differences should be smaller than variance accounted for by within-country differences:* The overall variance due to between-group differences contains variance based on C, C×D, and I:D×Ceffects. The overall variance due to within-group differences contains S:C and S:C×D effects. When the variation due to between-group differences is smaller than that due to within-group differences, it refers that cross-cultural differences are less critical than within-country differences. This result enhances cross-cultural applicability of the scale.

(2) *Within each dimension of the attitudes towards mathematics scale, the item facet (I:D) should be relatively small. The cross-cultural reliability coefficient (RC) should be at least 0.7 for the generalizability of the items:* Some variation of responses across items in the same dimension (I:D) is to be expected. Otherwise, if this variance is zero, it refers that the items are redundant or overlapped. In measurement tool development, it is important to contain items that tap different facets of dimension while excluding overlapping items. However, too much variation is undesirable either, as this might imply that the measurement instrument is not well expressed and functionalized. A relatively small variation due to (I:D) suggests that the items have internal consistency reliability. Also, the size of RC (>0,70) provides support cross-cultural applicability and items related to dimension have cross-national reliability (Nunnally & Bernstein, 1994).

(3) *The variance accounted for by I:D x C should be relatively small:* For each dimension, one would expect diversity in responses to items because of country differences. If the response patterns to items were to vary too much across countries, then this implies that the scale items are country-specific, and not cross-cultural applicable.

(4) *The variance accounted for by the dimension facet should be smaller than the variance accounted for by the person and country facets:* Too high of variance due to

D facet is undesirable. Each dimension relates to a different part of the attitudes towards mathematics concept. Thus, the variance accounted by the D facet is not expected to be zero. Furthermore, as items related to different dimensions of a construct vary, students' responses to the items of those dimensions are also expected to diverge. As student is a differentiation factor, the variance accounted for by the interaction of dimensions and students within countries (i.e., S:C×D) should be meaning. A significant S:C×D interaction would also support the discriminant validity among dimensions, thus enhancing the generalizability of the measurement tool.

(5) *If a scale is actually multi-dimensional, the generalizability coefficient (GD) for the whole measurement tool would be significantly smaller than the GD computed for each dimension separately:* The smaller G-coefficient for multidimensional scales is due to the dimension effect. This also supports the discriminant validity of scale dimensions

Table 1. Source of variation and its impact on generalizability.

Source of variance	Percent of variance	Effect on generalizability
Country (C)	$\sigma^2_{C}/\sigma^2_{total}$	• High variation implies that countries differ with respect to construct mean. • Variation associated with this source or factor is desirable; such factors are called differentiation factors or objects of measurement. • Compared to other factors. greater variation in this factor increases scale generalizability.
Student within country (S:C)	$\sigma^2_{S:C}/\sigma^2_{total}$	• High variation implies there is variation in student responses within countries • This factor is also a differentiation factor as variation due to this source is desirable. • Compared to other factors. greater variation in this factor increases scale generalizability.
Dimension (D)	$\sigma^2_{D}/\sigma^2_{total}$	• Lack of variation indicates that scale dimensions are overlapping; scale dimensions do not have discriminant validity. • High variation indicates the dimensions may be representing different constructs. but not the same construct. • This factor is a generalization factor; it is important to control variation due to this factor.
Item within dimension (I:D)	$\sigma^2_{I:D}/\sigma^2_{total}$	• Lack of variation indicates item redundancy • High variation indicates that a) the construct is poorly defined, or the measure is underdeveloped and • b) the measurement error is relatively high. • This factor is called the generalization factor; controlling variation due to this factor enhances generalizability.
Country by dimension (C x D)	$\sigma^2_{CxD}/\sigma^2_{total}$	• Low variation implies the pattern of responses to scale dimensions are the same across countries • High variation decreases scale generalizability
Country by item within dimension (C x I:D)	$\sigma^2_{CxI:D}/\sigma^2_{total}$	• Low variation implies responses to items of the same scale dimension do not vary across countries. If • this was true. then scale items are not country specific—a desirable outcome • High variation implies items of the same scale dimension are viewed differently in different countries—an undesirable outcome • Controlling for variation due to this interaction enhances

		scale generalizability
Student within country by dimension (S:C x D)	$\sigma^2_{\mathrm{S:CxD}}/\sigma^2_{\mathrm{total}}$	• Variation due to this factor is expected. as each scale dimension measures a different aspect of the underlying construct. and as student responses are expected to vary across dimensions • As differentiation factors enhance scale generalizability. variation due to this factor is desirable
Other interactions and error	$\sigma^2_{\mathrm{e}}/\sigma^2_{\mathrm{total}}$	• Low variation enhances scale generalizability

Source: Durvasula & Lysonski, 2016; Durvasula et al., 2006; Malhotra & Sharma, 2008

Method

Datasets/Participants

For illustration G-theory application in cross-cultural validation and generalizability of measures, Student Questionnaire dataset in PISA 2012 was used. Three countries were separated from this dataset: Türkiye, Finland, and the USA. These three countries were chosen to provide a high level of cross-cultural diversity. Then, randomly 1000 students were selected from Türkiye, Finland, and the USA, respectively.

Data Collection Tools

In the PISA 2012 Student Questionnaire, ten scales about attitudes towards mathematics were constructed using 67 items. Attitudes towards mathematics received considerable attention in the PISA 2012 Student Questionnaire (OECD, 2014, p. 320). Hence, in this research, eight out of these ten scales (with 49 items) were used and are summarized in Table 2.

Table 2. Attitudes towards mathematics scales.

Scale	Scale label	Number of items
INTMAT	Mathematics interest	4
INSTMOT	Instrumental motivation for mathematics	4
MATHEFF	Mathematics self-efficacy	8
ANXMAT	Mathematics anxiety	5
FAILMAT	Attributions to failure in mathematics	6
MATWKETH	Mathematics work ethic	9
MATINTFC	Mathematics intentions	5
MATBEH	Mathematics behavior	8

It was assumed that these scales were subscales of a structure called attitudes towards mathematics, although no such structure was actually defined. Therefore, dimensionality became a source of variance for G-study. Then To test this hypothetical latent model, CFA was conducted with Mplus 7 (Muthén & Muthén, 2012). Results from CFA revealed significant factor loadings for all items, ranging 0.27-0.95, and correlations among dimensions were significantly ranged from -0.42 to 0.74 (Table 3). Three fit indices were used to assess model fit: Tucker–Lewis index (TLI), the comparative fit index (CFI), and the root mean squared error of approximation (RMSEA). Çokluk et al. (2010, pp.271-272) suggested that an RMSEA value of < 0.05 indicates a close fit, and that < 0.08 suggests a reasonable model–data fit, and also recommended that TLI and CFI> 0.90 indicate an acceptable fit. According to fit indices (TLI=0.89; CFI=0.90; RMSEA=0.07), model-data fit only marginally supported. Table 3 shows the correlations between attitudes towards mathematics scales and scale reliabilities (Cronbach alpha) in countries.

Table 3. Correlations between attitudes towards mathematics scales and scale reliabilities in countries.

Scale	1	2	3	4	5	6	7	8
1. INTMAT	1.00							
2. INSTMOT	0.74	1.00						
3. MATHEFF	0.53	0.46	1.00					
4. ANXMAT	-0.42	-0.33	-0.44	1.00				
5. FAILMAT	-0.36	-0.32	-0.35	0.65	1.00			
6. MATWKETH	0.60	0.56	0.46	-0.25	-0.25	1.00		
7. MATINTFC	0.46	0.42	0.34	-0.26	-0.23	0.27	1.00	
8. MATBEH	0.58	0.39	0.39	-0.05	-0.04	0.51	0.26	1.00
Türkiye	0.89	0.87	0.71	0.82	0.66	0.91	0.77	0.80
Finland	0.90	0.89	0.85	0.82	0.68	0.88	0.83	0.72
USA	0.91	0.91	0.85	0.88	0.73	0.88	0.76	0.80

This is an unbalanced study since every dimension had a different number of items. Then, G-theory multi-facet analysis S:C×I:D to examine the cross-cultural applicability of a multidimensional structure was performed using G String IV program (Block & Norman, 2012), based urGENOVA by Brennan (2001b). The analysis of S:C×I:D design generated variance components for country (C), dimension (D), student within country (S:C), item within dimension (I:D), country × dimension interaction (C×D), country × item interaction (I:D×C), student × dimension interaction (S:C×D), and error. Dimension-level analysis of S:C×I design generated variance components for C, S:C, C×I interaction, and error.

Like in ANOVA, overall score variance is divided components of variance into individual facets and their interactions. The sources of variance represent which facets or interactions contribute to considerable amounts of error variance. The variance components are also used to calculate generalizability coefficients. The formula for computing the G-coefficient is as follows (Crocker & Algina, 1986; Shavelson &Webb, 1991):

$$G = \frac{\sigma^2_{tru\ score}}{\sigma^2_{true\ score}+\sigma^2_{error\ score}} \qquad (2)$$

The components that account to error variance differ if absolute ($\sigma^2_{absolute}$) or relative ($\sigma^2_{relative}$) is predicted. The $\sigma^2_{absolute}$ is more convenient when the judgment will be absolute against determined standards. A given instance in the study by Durvasula et al. (2006), a country's point on a promotional measure may be regarded acceptable when the point overreaches some preconcerted standard. The judgment is absolute because of not depending on an ordering of other countries. Conversely, the country's point may be regarded acceptable when it passed another country's point. In this situation, the judgment depends on an ordering, so the $\sigma^2_{relative}$ is more convenient. For this reason, generalizability and reliability coefficients were calculated for the findings in the present paper. The following equations display how to calculate the generalizability coefficients:

$$GC_1 = \frac{\sigma^2_{true\ score}}{\sigma^2_{true\ score}+\left(\frac{\sigma^2_{CxD}}{D}+\frac{\sigma^2_{I:DxC}}{IxD}+\frac{\sigma^2_{S:CxD}}{D}+\frac{\sigma^2_{Error}}{IxD}\right)} \qquad (3)$$

$$\mathrm{GC}_2 = \frac{\sigma^2_{true\ score}}{\sigma^2_{true\ score}+(\frac{\sigma^2_{IxC}}{I}+\ \frac{\sigma^2_{Error}}{I})} \qquad (4)$$

where the $\sigma^2_{true\ score} = \sigma^2_C + \sigma^2_{P:C}$, I is the number of items, and D is the number of dimensions. For cross-cultural applicability, item and dimension are the generalization facets, whereas country and students are the differentiation facets. Therefore, whereas GC$_1$ is computed for the S:C×I:D design (Durvasula et al., 2006), then GC$_2$ is computed for the S:C×I design or dimension-level analysis (Malhotra & Sharma, 2008).

Estimates of variance components for each dimension could also be used to obtain measurement consistency. So, the desirable variance comes from the S:C facet, but the undesirable variance comes from the error. Malhotra and Sharma (2008) have stated that “The error component excludes variation due to items as CTT assumes parallel measures and its effect is considered constant across all subjects. In other words, one is interested in determining the extent to which subjects’ scores can be generalized across items” (p. 651). The cross-cultural reliability coefficient (RC) is then given by

$$\mathrm{RC} = \frac{\sigma^2_{S:C}}{\sigma^2_{S:C}+\ \frac{\sigma^2_{Error}}{I}} \qquad (5)$$

Results

Table 4 shows multi-facet analysis result of the S:C×I:D design result for multidimensional attitudes towards mathematics construct, and Table 5 displays the analysis result of the S:C×I design for the dimension-level.

Table 4. Estimate variance components of the S:C× I:D design

Source of variance	Variance component	Percentage of total variance	G coefficient
C	0.01	1.46	0.70
S:C	0.07	7.91	
D	0.12	14.71	
I:D	0.05	5.53	
C x D	0.01	1.31	
I:D x C	0.02	2.13	
S:C x D	0.25	30.21	
Error	0.30	36.73	
Total	0.84	100	

Note: C=Country, S=Student, D=Dimension, I=Item

As seen in the Table 4, variance accounted for by between-country, or cross-cultural, differences that contain C, C×D, and I:D×C of variance components is 4.90% (1.46% + 1.31% + 2.13%), while the variance accounted for by within-country differences which contains S:C and S:C×D of variance components is larger at 38.12% (7.91% + 30.21%). This result implies that variance from personal (within-country) differences is about eight times the size of variance from between-country differences. Accordingly, variety in student responses within the three countries is greater than diversity across the countries. These findings confirmed the first research expectations and generalizability of attitudes towards mathematics construct measure.

Table 5. Estimate variance components of the S:C×I design for the dimension-level

Dimension	Variance component					Coefficient	
	C	S:C	I	C x I	Error	GC	RC
INTMAT	0.05	0.56	0.01	0.01	0.24	0.91	0.90
INSTMOT	0.00	0.43	0.00	0.01	0.24	0.87	0.87
MATHEFF	0.02	0.26	0.05	0.02	0.42	0.79	0.78
ANXMAT	0.03	0.39	0.06	0.02	0.37	0.85	0.82
FAILMAT	0.04	0.25	0.03	0.03	0.63	0.72	0.71
MATWKETH	0.04	0.30	0.02	0.02	0.31	0.86	0.85
MATINTFC	0.00	0.10	0.01	0.00	0.14	0.79	0.79
MATBEH	0.04	0.22	0.09	0.04	0.43	0.82	0.79

Note: C=Country, S=Student, D=Dimension, I=Item

From Table 4, the variance accounted for by the item facet (I:D) (one of the undesirable sources of variance) is relatively small (5.53%) but different from zero. Also, variance accounted for by item in Tablo 5 is insignificant for all dimensions. These results signify that the items for each dimension are nearly homogeneous. From Table 5, the values of RC are above 0.7 for all eight dimensions/scales, which indicates that the items under different dimensions demonstrate high cross-cultural internal consistency reliability with confirming the second research expectation. The contribution of the I:D×C variance component to overall variance is relatively small at 2.13%. Also, variance accounted for by C×I interaction is small for all dimensions in Table 5. This finding refers that the pattern of students' responses is the same across the three countries, and that the items are not country specific, which confirms the third research expectation.

The dimension facet accounts for 14.71% of the overall variance. This rate is somewhat high, but not surprising. A priori, one would expect that the variance introduced by the D facet must be smaller compared to the variance from the S and C facets for enhancing cross-cultural generalizability or applicability. However, the findings show that the variance from the differentiation facets (S and C) is significantly smaller than the variance from the D facet. The multidimensional structure analyzed in this study does not belong to a scale developed with strong structural evidence, it was just assumed that it was the case by the researcher (which the model-data fit was marginally supported as described in the method). Accordingly, it actually shows that the G- theory is also useful to look for evidence based on the structural validity of the measurements. One would expect that another priori, the variance from D facet not to be too small or zero. This is because the trivial contribution from the D facet to overall variance refers to the lack of distinction or little difference between the dimensions. The findings display the opposite case in the present study. Also, S:C×D interaction largely accounts for 30.21% of the overall variance, which refers that student evaluated diverse dimensions of the attitudes towards mathematics differently. There is a confirmation for discriminant validity between the eight dimensions of the attitudes towards mathematics scale with supporting the fourth research expectation.

The GC and RC estimates for all dimensions range from 0.71 to 0.91 (see Table 5), which are above 0.70 recommended by Rentz (1987). On that account, the items for eight dimensions of attitudes towards mathematics could be generalized across students and countries. The value of GC for the multidimensional construct, which is equals to 0.70 (see Table 4), is lower than that estimated from dimensions because of the considerable dimension effect. This confirms the fifth research expectation.

Conclusion and Discussion

Large-scale assessments, such as PISA, TIMSS, have an important place in educational science research. Applying a measurement instrument developed in a specific country to other country/countries raise a critical and important question of interest in especially cross-cultural studies. Are the operational definition and notional sense of a specific construct of the administered measurement instruments in these assessments the same between countries of attended? Do the measures have similar levels of reliability or validity in all countries? CFA, whose framework suggested by Steenkamp and Baumgartner (1998), has been the most common way to answer these questions and the cross-cultural generalizability or applicability of measurement tools. Although CFA is a sophisticated technique for examining various equivalence types (structural, metric, scalar etc.), the technique has some limitations. Because of based the CTT, when a measurement tool is not invariant between countries, what factors contribute to the error variance is unclear. CFA is not able to distinguish amongst possible sources of response variance. Also, the object of measurement is assumed to be person in CFA. Therefore, CFA becomes useless when the object of measurement is not person but country, rater, dimension, or item. In the literature, some researchers in management, business or marketing area (Sharma & Weathers, 2003; Durvasula et al.,1993, 2006; Malhora & Sharma, 2008; Durvasula & Lysonski, 2016) outlined the usefulness of G-theory in assessing cross-cultural applicability of measurement instruments. Hence, a fundamental focus of this study is to examine measurement comparability or cross-cultural applicability for different countries on an international assessment using G-theory in educational science studies. With multi-facet design, the contribution of dimensionality to error variance is examined, too. For illustration purposes, eight scales from the PISA 2012 student questionnaire dataset related to attitudes towards mathematics was used. Türkiye, Finland, and the USA were chosen from this dataset to provide a high level of cross-cultural diversity. For the analysis, G String IV program was used because of the unbalanced design.

Almost all of the results supported all research expectations. According to the parameters of the G-theory, it can be rightly deduced cross-nationally applicability of the attitudes towards mathematics scales can be from these research findings. Only attributions to failure in mathematics scale possess moderate size GC and RC estimations, all other scales have a high level of coefficients. The variance component associated with the country facet is insignificant for either the multidimensional scale or each dimension. This implies that countries have similarities in construct mean. In comparison, a large variance component associated with the student within-country facet (S:C) implies that there is divergence in student responses within the various countries to scale dimensions. But multidimensional measures of attitudes towards mathematics possesses limit level of cross-cultural applicability because of the undesirable level of dimension component. As noted, this supports the that the multidimensional structure does not belong to a scale developed with strong structural evidence.

In conclusion, this study ensures an easy procedure to apply to social and educational science researchers and practitioners for cross-cultural scale applicability and validation by presenting the application of G-theory. This study does not say or defend that the use of CFA is inadequate in terms of measurement equivalence. This is only to show that multi-facet analysis can be used for cross-cultural studies and has many advantages.

Limitations and Suggestions for Future Research

A limitation of the study is the conducting of a multidimensional artificial structure by using the PISA 2012 dataset since any data belonging to a multidimensional test administration could not be found. Another limitation is that only three countries were compared. Therefore, it is recommended to conduct a similar study using the validated measurement results for future studies. This study only focuses on multi-facet analysis. Durvasula et al. (1993, 2006) has claimed that a small variance of items within dimensions by country interaction is similar to metric invariance in CFA procedures. Hence, as another research topic, a comparison between CFA and G-theory can be designed by invariance types. How between-dimension correlation affects the parameters of the G-theory or cross-cultural applicability can be examined also.

References

Block, R., & Norman, G. (2012). G String IV (Version 6.2.1.2). [Software]. Available from http://www.papaworx.com/

Brennan, R. L. (2001a). *Generalizability theory.* Springer

Brennan, R. L. (2001b). *Manual for urGENOVA.* Iowa City, IA: Iowa Testing Programs, University of Iowa.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Wadsworth Thomson Learning.

Cronbach, L.J., Rajarathnam, N. & Gleser, G.C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology 16*(2), 137–163.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları* [Multivariate statistics for social sciences: SPSS and LISREL applications]. Pegem Akademi

Durvasula, S., Andrews, J. C., Lysonski, S., & Netemeyer, R. G. (1993). Assessing the cross-national applicability of consumer behavior models: A model of attitude toward advertising in general. *Journal of Consumer Research, 19*, 626–636.

Durvasula, S., Netemeyer, R. G., Andrews, J. C., & Lysonski, S. (2006). Examining the cross-national applicability of multi-item, multi-dimensional measures using generalizability theory. *Journal of International Business Studies, 37(*4), 469–483.

Durvasula, S. & Lysonski, S. (2016) Finding cross-national consistency: Use of G-theory to validate acculturation to global consumer culture measure. *Journal of Global Marketing*, *29*(2), 57-70.

Eisend, M. (2009). A cross-cultural generalizability study of consumers' acceptance of product placements in movies. *Journal of Current Issues & Research in Advertising, 31*(1), 15-25.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117 – 144.

Johansson, S. (2016). International large-scale assessments: what uses, what consequences?, *Educational Research, 58*(2), 139-148

Malhotra, M. K., & Sharma, S. (2008). Measurement equivalence using generalizability theory: An examination of manufacturing flexibility dimensions. *Decision Sciences, 39*(4), 643–669.

Matsumoto, D., & Van de Vijver, F. J. R. (Eds.). (2010). *Cross-cultural research methods in psychology*. Cambridge University Press.

Muthen, L., & Muthen, M. (2012). Mplus Software (Version 7). [Software]. Available from https://www.statmodel.com/

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.

OECD (2014). *PISA 2012 technical report.* OECD Publishing.

Rentz, J. O. (1987). Generalizability theory: A comprehensive method for assessing and improving the dependability of marketing measures. *Journal of Marketing Research*, *24*(1), 19–28.

Sharma, S., & Weathers, D. (2003). Assessing generalizability of scales used in cross-national research. *International Journal of Research in Marketing*, *20*(3), 287–95.

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability theory: A primer.* Sage Publications

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90.

Van de Vijver, F.J.R. & Leung, K. (1997). Methods and data analysis of comparative research. In J.W. Berry, Y.P. Poortinga and J. Pandey (eds.), *Handbook of Cross-Cultural Psychology*, *Volume One: Theory and Method* (pp. 247–300). Allyn & Bacon.

Van de Vijver, F.J.R. & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumo and F. J. R. Van de Vijver (eds.), *Cross-cultural research methods in psychology* (pp. 17-45), Cambridge University Press.

Van de Vijver, F. J. R., Leung, K., Fetvadjiev, V. H., He, J. & Fontaine, J. R. (2021). *Methods and data analysis for cross-cultural research* (Second edition). Cambridge University Press.