Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Using G-theory in the development of performance assessment of the socio-emotional domain of children

M.R. Nor Mashitah [1, 2, *], M.N. Mariani [1], Shahrir Jamaluddin [1], Mohd Nazri Abdul Rahman [1]

[1]Department of Educational Psychology and Counseling, Faculty of Education, University of Malaya, Kuala Lumpur, Malaysia
[2]Department of Early Childhood and Education, Faculty of Education and Human Development, Sultan Idris Education University, Perak, Malaysia

ARTICLE INFO

Article history:
Received 11 November 2016
Received in revised form
17 January 2017
Accepted 21 January 2017

Keywords:
Socio-emotional domain performance-based assessment
G-Theory

ABSTRACT

Performance assessment socio-emotional domain, through the ability of children in the context of the performance criteria. This study, investigates potential applications of Generalizability theory (G-theory) in the development of such a performance-based assessment procedure. 77 kindergarten children were assessed as participants in this study. Firstly, analysis of variance showed that nested rater variance component in person and item (r:pi) component accounted for the highest percentage of the total variance, 0.24942; 42.2% and the smallest, variance of items 0.04232; 7.2%. Secondly, through analysis in G-study, 94% of the overall variance can be explained by the design. Next, based on optimization analysis in D-study that the overall absolute Coefficient G reading remains at 0.97, which was an acceptable value. Lastly, for reliability test from G-facets analysis, the overall cognitive domain reliability was recorded at 0.96 as the reliability of the 38 items was ranging to 0.96. This study base on Theory-G had an impact on minimizing the error of measurement and determining the appropriateness use of items in the administration of the assessment.

1. Introduction

Task completion in an actual context includes performance based assessment. The ability to complete a task demonstrates the real capability of the children. The actual abilities, it is in line with the authentic terminology that is authentic assessment (actual assessment) or authentic performance (actual performance) (Nancy, 2001). If performance assessment or authentic assessment is used to understand how children relate or apply what they have learned, the learning experience provided must be authentic and meaningful as well. When children are related to authentic learning, they are given the opportunity to link new information with the existing information while solving problems.

To clarify the relevance of authentic learning and actual abilities of children, it is appropriate to refer to Kleinert et al. (2002) who stated the objectives of this approach is to allow children to show how they use what they know to represent learning in the form of product or performance. In other words, by authentic learning, it has stimulated children to show their knowledge or true feelings of themselves. According Wehlage et al. (1996), authentic learning fosters knowledge construction and focuses on higher-order thinking. The aim is to enhance knowledge level and construct new knowledge. To understand how children relate or apply what has been learned, the learning experience provided must be authentic and meaningful also. When a child is associated with authentic learning, they are given the opportunity to connect new information with existing information while solving the problem. Therefore, for children, opportunities provided through a variety of activities during assessment is intended to observe the level of existing knowledge and new experiences as well as new knowledge when aid granted during activity.

In particular, researchers have been looking at the issue of diversity in assessment tasks and consideration rater as a source of measurement error in the performance test. It is considered that the procedure will lead to a reduction in error associated with the use of human evaluators. This can happen, but it is not always certain; judgment is required to develop rules or protocols for

* Corresponding Author.
Email Address: normashitah1604@gmail.com (M. R. Nor Mashitah)
https://doi.org/10.21833/ijaas.2017.03.022
2313-626X/

computerized scoring algorithm. Decisions made by different experts and/or types of experts may lead to different computerized scoring algorithm (Clauser et al., 1999). This study aimed to investigates potential applications of Generalizability theory (G-theory) in the development of children performance assessment for socio-emotional domain (Cardinet et al., 2009).

2. Literature review

Previous studies of performance-based assessment using instrument is to support children through evidence and proof obtained as well as to identify the strength and weaknesses (Gardner, 1993). When referring to the first purpose of using performance-based assessment on children, it is assure that this assessment is a good tool to assess the progress of development of children as performance-based assessment is designed to measure the actual performance of children or their assignments or activities related to learning. Using observations towards performance is closely related to or directly linked to the development of the achievement rate in children (Harrington et al., 1997). Secondly, performance-based assessment is integrated to teaching. Performances in activities are the natural learning outcome that is parallel to the curriculum and teaching which cannot be separated. Hills (1993) elaborated while using performance-based assessment, teachers have to know the suitability of its design, the relationship as a mean of testing, interpret the results of assessment to understand the progress of the children, plan further lessons and deliver results to parents and administrators.

Dependability refers to accuracy in generalizing scores obtained from respondents in a test to average score obtained by students in various situations (Shavelson and Webb, 1991). In this research context, dependability is the index obtained in a test analysis based on different individual and item.

This research is about G-Study to identify various variance resources which might be in an assessment by estimating the variance component that is contributed by each of it. It is carried out to evaluate the measurement of dependability that is done to the variance which can be considered in the future measurement. This research is focus on D-Study to put forward reliability coefficient as generalizability coefficient covering variance towards error resources. D study is also able to differentiate between the relative decision and the ultimate decision. By using the information which has been collected through G study, D-study can design a better and more suitable measurement application for a measurement and assessment suggested (Shavelson and Web, 1991).

Based on the findings of a study D-measurement protocols for different scenarios, to achieve a good balance between reliability and cost efficiency, past research recommend using two independent raters for each class of kindergarten (Dezhi et al., 2014). Furthermore, the use of G-theory in the context of this new chapter in the evaluation of the quality of child cares programs.

G-Theory produced a more integrated approach to assess reliability which has been carried out whether for the purpose of making relative decision (norm-reference test) or actual decision (criteria reference test). Relative decision is based on individual's place in a group compared to actual decision. Actual decision is based on actual score without any comparison with other individuals score in the group (Ary et al., 1996) and decisions at the individual student level (Fan and Hansmann, 2015). G-Theory does not make assumption regarding comparison of error resources, but estimate simultaneously the variety of error resources including interaction between those errors. To compare the impact of raters and tasks on reliability, they computed average reliability due to raters and known as "score reliability" (Brennan, 2000).

Previous research also show, the analyzed how variation in facet number affects reliability with the testing of reliability in generalizability theory by using different designs (Büyükkidik and Anil, 2015). It is a generalization of the classical reliability theory, which examines the relative contributions of the main variables of interest, the performance of subjects, versus error variance. In theory G, various sources of error contribute to inaccuracy of measurement will be explored. G theory is an effective tool in assess the methodological quality of assessment methods and improve accuracy (Ralph and Geoffrey, 2012).

3. This study

This study emphasized performance-based assessment towards physical socio-emotional in a fun learning environment which involves learning activities with teachers in the playschool. To assess is to collect information. Observation method is used to collect information and evidences. Observation means children's behavior is under scrutiny. This approach can be used without the consciousness of the children that they are being observed. This study used the role of the Rater, which is the teachers themselves observe the children. Every child will be evaluated by raters. Generalizability theory or G theory is particularly well suited to addressing such matters in that it enables an investigator to quantify and distinguish the sources of inconsistencies in observed scores that arise, or could arise, over replications of a measurement procedure (Brennan, 2010).

The broad research question that guided this study was: a) what is the contribution of facet towards variance resource according to the Generalizability Theory, (b) what is the score coefficient value of children's performance according to the G-Study, (c) what is the best optimization value towards facet in order to increase the value of coefficient G by using D-study, and (d) the reliability

score for each item in the performance-based assessment in G-facets analysis.

4. Methodology

Research design of this study is in the form of survey and analyzed data in quantitative method. Computerized scoring procedures for performance assessment are currently receiving considerable attention (Bennett, 1999; Bennett et al., 2000; Brennan, 2000). This study is a descriptive research in order to collect feedback from respondents as well as to survey error resources in measurement. Research design is as Table 1.

Dependability of test score will be used Two facet (r:pi) partially Nested Random Design. Data will be analyzed using EduG software in order to get result for G study and D study. Design model of two facet (r:pi) partially Nested Random Design is as shown in Fig. 1 and Fig. 2.

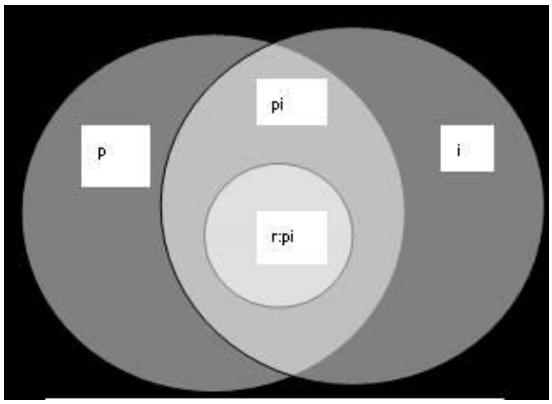

Fig. 1: Variance resources

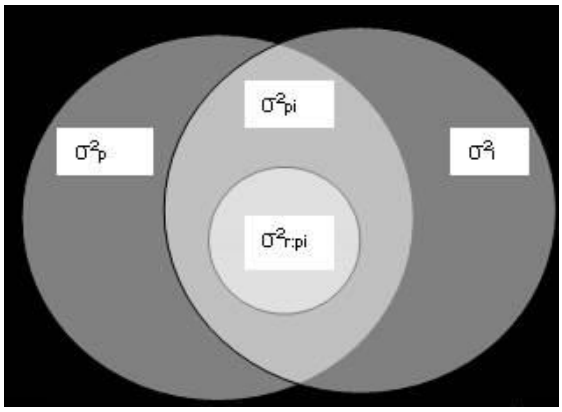

Fig. 2: Variance components

Fig. 1 shows Venn graph for the research design of this study, that is Two facet (r:pi) partially Nested Random Design. Fig. 2 is component of variance resources. The p circle represent children (person) being evaluated in the domain of socio-emotional development. However, the item circle, i represent item of the socio-emotional development domain which is tested on children. This item is made up of item which requires children to show response of their ability in doing it. The person circle, p intersects with the item circle, i produced interaction between people and item, that is the pi interaction. pi interaction shows how children give response towards item which is being tested in the assessment. Following that, in the intersect part between the p and i circle, nested circle is the rater, r. This shows that different rater will evaluates the children's performance, yet item being tested is the same.

Table 1: Research design

Research Design	Data Collection Method	Respondent	Data Resource
Quantitative research	NOaMA Instrument	18 raters who give score to the performance of 77 children	Performance score

In this study, person (p) is the object of measurement. Two facet involved is the nested rater (r) and item (i) in children as well as item p/ri. Observation design is r:pi. All measurement object children and facet are infinite random because the population of inspector and student are infinite, also having variability with universe set.

Table 2 shows variance resources in this study. Based on the research design, two facet (r:pi) partially Nested Random Design, it has produced 4 variance resources, that is person (p), item (i), rater nested in children and item (r:pi) as well as interaction between person item (pi) and residual (e).

4.1. Sample

Based on the 77 children who were enrolled in the registered playschools, sample selection is based on stratified random method. A total of 77 children as research sample represented the population being studied. Performance-based assessment were carried out among 77 children and were given scores by two different raters (teacher) from 9 playschools, that is all together 18 raters and identified as rater 1 and 2.

Table 2: Variance resources for two facet partially nested (r:pi) design

Variability Resources	Variance Note
Person (p)	$\sigma^2(p)$
Item (i)	$\sigma^2(i)$
Rater nested in pi	$\sigma^2(r:pi)$
Interaction pi, e	σ^2pi, e

4.2. Research instrument

NOaMA assessment is a learning assessment approach and children development in this study have been re-designed in year 2013 to include scoring procedures in Likert scale (5 points). This instrument was re-designed to comply with the assessment concept of National Early Childhood Care and Education Policy,

This instrument reflects the overall skill at the age group which requires children to relate with the learning and development domain. The socio-emotional domain contains performance's item that require children to perform a task. Activities

prepared will translate such performance items. The socio-emotional domain contains a number of 38 items.

The data has been analyzed by using EduG is able to estimate every variance component and determine the dependability score in a test. There are various designs that can be analyzed using EduG according to the desired facet. In the research carried out, researcher used the Two-Facet Partially Nested Design. Analysis outcome of EduG have produced two types of research, that is G-study (Generalizability studies) and D study (Decision Studies). G-study is able to identify variance resources and variance magnitude, while D-study is able to determine coefficient G as well as the design suitable to the number of item in a particular test.

5. Results

5.1. The contribution of facet towards variance resource according to the Generalizability Theory

From analysis, variance component which contributed to the dependability of test is shown in Table 3.

Table 3: Variance component of performance based assessment on socio-emotional development domain

Facet	Df	Total Chi Square	Mean Chi Square	Variant Component	% Variant
Children, K	76	1190.97369	15.67071	0.19502	33.0
P:KI	3003	749.00000	0.24942	0.24942	42.2
Item	38	265.08425	6.97590	0.04232	7.2
KI	2888	1325.40293	0.10476	0.10476	17.7
Total	6005	3530.46087			100
Coefficient G_absolute	0.97				

Table 3 shows the variant component of each facet which contributed to the difference of children's score evaluated by the rater in the assessment of socio-emotional development domain performance. The variant analysis shows that the variant component of nested raters in person and item (pi) shows the highest value of variant component, which is 42.2% followed by variant is 33% the person variant component (p). Next, 17.7% is the interaction among person and items (pi) and the smallest component is the variant of item (i) which is 7.2%.

Based on the analysis, the variant component of nested rater in person and item (r:pi) indicates the highest value shows (σpki = 0.24942; 42.2% from the overall total of variant component). This shows that there are differences between raters in giving scores to the children. This is because raters had understood that the scoring based on rubrics and all raters have dissimilar consistency, while giving scores for the evaluation of socio emotional development domain.

Through the analysis, it was found that variant component of person, (p) shows the highest value of variant component, (σ^2p = 0.19502; 33.0% from the overall total variant component). This shows that the children abilities are significantly different and it means that the children who participate have different abilities.

Next, variant component that shows average reading is the interaction among person and item (pi) which is (σpi = 0.10476; 17.7% from the overall total of variant component). This shows that there is average difference among the children in giving response on the tested items.

The smallest variant component is the variant component of items (i) which indicates the lowest value of component (σ^2i = 0.04232; 7.2% from the overall total variant component). This shows that children dependability in the test is not influenced by items. The lowest percentage for item component shows that tested items in the evaluation is different in terms of difficulties. The different in level of difficulties influence the performance showed by the children.

5.2. The score coefficient value of children's performance according to the G-Study

Relative coefficient G (0.97) and absolute coefficient G (0.97) in Table 4 showed value beyond the accepted conventional value, 0.8. Research design is good to analyze children's dependability score because coefficient G value beyond conventional value. Absolute coefficient G is considered as this research aimed to evaluate children's dependability score individually based on the contribution of variant component in different raters.

Through analysis, 94% of the findings from children's score are attributable to the universe score. This means that 94% of the overall research can be explained. However, only 6% of finding score is attributable to random impacts which are not identifiable.

This design produced reliability measurement or dependable measurement and the advantages of using Generalizability Theory analyses to examine score reliability (Arterberry et al., 2014). It is also can be interpreted as 94% of the factors that contributed to the children's variance score can be explained, while 6% contributing factors found from error resources which are not identifiable. Findings also show that standard error related to children's decision score is small while absolute standard error is 0.08348. Standard error shows value that is smaller than the estimated standard deviation 0.44161 for true score dispersion.

Table 4: G- study

Source Of Variance	Differ- Entiation Variance	Source Of Variance	Absolute Error Variance	% Absolute
K	0.19502		.....	
	.....	P :KI	0.00320	45.9
	.....	I	0.00109	15.6
	.....	KI	0.00269	38.5
Sum of variances	0.19502		0.00697	100%
Standard deviation	0.44161		Absolute SE: 0.08348	
Coef_G relative	0.97			
Coef_G absolute	0.97			

5.3. Best optimization value towards facet in order to increase the value of coefficient G by using D-study

In D-study, the relative coefficient G ($\hat{E}p^2$) displays different level of relative error variance. In D-study, absolute coefficient G phi (Φ) shows degree of difference in absolute error variance. Table 5 shows the difference of reliability value or coefficient G when number of children and rater increase or modified.

In this research, the absolute coefficient G phi (Φ) will only be taken into account because this research is to examine error variance towards children's score evaluated by two different raters in the performance based assessment in socio-emotional development domain in playschools. This research also compares score given by two raters of different playschools.

Based on Table 5, it is found that number of children that are suitable to be evaluated in the assessment is 77 by taking into account the number of raters remained at 2 person. With reserves of 77, the Coef_G absolute phi (Φ) remained at 0.97 which is a high value and it is accepted. Coef_G absolute value of phi (Φ) exceeds the accepted conventional value 0.8. The decision to choose the number of children that are suitable for assessment is based on the consideration of factors such as time, cost, logistics and others. This means that if the number of children which were maintained at 77 children; it is accepted and sufficient to deal with restrictions on time, cost logistics and others.

Therefore, for this study, researcher suggested number of children to be 77 children and 2 raters in the performance based assessment in the socio-emotional development domain is maintained for the value of coef_G absolute phi (Φ) or high reliability parallel with these findings.

5.4. Reliability score in the performance-based assessment in G-facets analysis

G-Facets Analysis is carried out to identify the contribution of each item to be tested in the performance-based assessment of the value of the coefficient G or reliability. This analysis estimates the coefficient G adequate for each item tested.

Table 6 shows the relative and absolute value of the coefficient G for each item tested. Generally, all items are functioning well because the value of coefficient G is greater than 0.8. Among these items, item 16 is seen as an item that contributed the largest error in the scoring to children. Item 16 can be said to represent an item which has a high difficulty level or testing children in achieving high level of performance. However, a conclusion can be made that these items are consistent as performance assessment items used to evaluate children. So, these items should be retained and can be used as a test set for children performance-based assessment bank item in socio-emotional development domain.

Table 5: The variances component of D Study based on the modification number of person and raters

Amount	G-Study	Opt 1	Opt 2	Opt3	Opt 4	Opt 5
Children (K)	77	100	120	140	160	180
Raters (P:KI)	2	4	4	3	2	2
Coef_G relative ($\hat{E}p^2$)	0.97	0.98	0.98	0.98	0.97	0.97
Coef_G absolute (Φ)	0.97	0.97	0.97	0.97	0.97	0.97

6. Discussion

Model design of this study is Two facet (r:pi) partially Nested Random Design, it has 4 variance resources, that is person (p), nested rater in person and item (r:pi), item (i), and interaction between person-item (ki) and residual (e).

The variant analysis shows that the variant component of nested raters in person and item (r:pi) shows the highest value of variant component, which is 42.2% followed by variant is 33% the person variant component (p). Next, 17.7% is the interaction among person and items (pi) and the smallest component is the variant of item (i) which is 7.2%. Based on the analysis, the variant component of nested rater in person and item (r:pi) indicates the highest value shows (σrpi = 0.24942; 42.2% from the overall total of variant component). This shows that there are differences between raters in giving scores to the children. This is because raters had understood that the scoring based on rubrics and all raters have dissimilar consistency while giving scores for the evaluation of socio-emotional development domain. Through the analysis, it was found that variant component of person, (p) shows the highest value of variant component, (σ^2p = 0.19502; 33.0% from the overall total variant component). This shows that the children abilities

are significantly different and it means that the children who participate have different abilities. Next, variant component that shows average reading is the interaction among person and item (pi) which is (σpi = 0.10476; 17.7% from the overall total of variant component). This shows that there is average difference among the children in giving response on the tested items. The smallest variant component is the variant component of items (i) which indicates the lowest value of component (σ^2i = 0.04232; 7.2% from the overall total variant component). This shows that children dependability in the test is not influenced by items. The lowest percentage for item component shows that tested items in the evaluation is different in terms of difficulties. The different in level of difficulties influence the performance showed by the children.

Table 6: G-Facets Analysis towards item (i)

No.	Coef_G Relative	Coef_G Absolute
1	0.97035	0.96497
2	0.96958	0.96407
3	0.96936	0.96379
4	0.97058	0.96521
5	0.97065	0.96554
6	0.96941	0.96386
7	0.96939	0.96375
8	0.96990	0.96433
9	0.96989	0.96439
10	0.96944	0.96385
11	0.97047	0.96514
12	0.96998	0.96452
13	0.97050	0.96508
14	0.96983	0.96443
15	0.96932	0.96373
16	0.97122	0.96646
17	0.96954	0.96407
18	0.97087	0.96579
19	0.96947	0.96395
20	0.96958	0.96413
21	0.96955	0.96422
22	0.96912	0.96352
23	0.97032	0.96489
24	0.96956	0.96405
25	0.96935	0.96378
26	0.96968	0.96424
27	0.96983	0.96434
28	0.96970	0.96416
29	0.96976	0.96427
30	0.96948	0.96396
31	0.97103	0.96631
32	0.97086	0.96613
33	0.96961	0.96415
34	0.97069	0.96569
35	0.97100	0.96617
36	0.96974	0.96507
37	0.96987	0.96489
38	0.97010	0.96461

Analysis based on Generalizability Theory by using EduG software is able to show variant component of every facet that contributed to the difference of children's score. G coefficient worth 0.97 is interpreted as 94% of the factors contributed to the children's score variance, while 6% of the contributing factors found from the error resources are not identifiable. The variant component of nested rater variant component in person and item (r:pi) shows the highest value of variant component 42.2% from the overall total of variant component. This shows differences between raters in giving scores to the children. This is because raters had understood that the scoring based on rubrics and all raters have dissimilar consistency while giving scores for the evaluation of socio emotional development domain. The person variant component (p) shows the highest reading 33.0% from the overall total of variant component. This shows that these respondents have huge differences in abilities.

Based on optimization analysis, it is suggested to remain the 77 children, with absolute Coef_G phi (Φ) which maintained at 0.97, that is a high value and accepted. This absolute Coef_G phi (Φ) value is beyond the accepted conventional value; that is 0.8. The decision to choose the number of children which is the most suitable for the assessment is made by consideration of factors such as time, cost, logistics and other. This means that if the number of children which were assessed remains at 77 children, it is accepted and sufficient to cope with the constraint of time, cost, logistics and others. Therefore, in this study, the researcher suggests the number of children to be remained at 77 children and rater 2 persons in the performance based assessment in the socio-emotional development domain in order to obtained high absolute Coef_G phi (Φ) value or high reliability value which parallel with the research findings.

Based on G-facets analysis, a conclusion can be made that these items are consistent as performance assessment items used to evaluate children. So, these items should be retained and can be used as a test set for children performance-based assessment bank item in socio-emotional development domain.

7. Conclusion

These findings lead to a number of implications in the construction of early learning standard instrument in early childhood development. Practically, it is difficult to build a truly fair and equitable item for all students who have different abilities. G-study and D-study according Generalizability Theory that have been carried out gives impacts in efforts to minimize the measurement error besides making wise decisions in number of item that is the most suitable to be administered in this assessment in the future. Items that functioned well can be included into the assessment item bank of socio-emotional development domain. Analysis of children's abilities by using rater assessment based on Generalizability Theory gives a different dimension compared to analysis based on CTT. By Generalizability Theory analysis, the contribution of each error in the measurement can be identified separately, which contrary to analysis of CTT, making analysis of Generalizability Theory a more precise and detailed. In assessing the ability of children, the set of assessment need to be implemented carefully after taking into account various factors that contribute to the result scores in the assessment. The constructor of the assessment item is responsible to ensure the constructed items show continuing consistency if

tested on other children and validated according to the needs and purpose the instrument is constructed. The existence of internal and external factors that may contribute to the variance of score should be controlled so that the reliability of findings and validity of the instrument can be improved. GT may explain the error components which become the contributing factor to the difference of assessment score. Analysis of socio-emotional development domain items based on the above theories has clarified directly or indirectly on the quality of the test and the improvements that need to be implemented to ensure that the instrument is truly able to meet the objectives of the measure.

Acknowledgment

Special appreciation to the Institute of Research Management and Monitoring (IPPP), University of Malaya Kuala Lumpur in allowing and giving postgraduate grant for us to conduct this study. Similarly, the cooperation of respondents and teachers from all kindergartens involved.

References

Arterberry BJ, Martens MP, Cadigan JM, and Rohrer D (2014). Application of generalizability theory to the big five inventory. Personality and Individual Differences, 69: 98-103.

Ary D, Jacobs LC, and Razavieh A (1996). Introduction to research in education. Harcourt Brace College Publishers, Florida, USA.

Bennett RE (1999). Using new technologies to improve assessment. Educational Measurement: Issues and Practice, 18(3): 5-12.

Bennett RE, Morley M, and Quardt D (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. Applied Psychological Measurement, 24(4): 294-309.

Brennan RL (2000). Performance assessments from the perspective of generalizability theory. Applied Psychological Measurement, 24(4): 339-353.

Brennan RL (2010). Generalizability theory and classical test theory. Applied Measurement in Education, 24(1): 1-21.

Büyükkidik S and Anil D (2015). Investigation of reliability in generalizability theory with different designs on performance-based assessment. Education and Science, 40(177): 285-296.

Cardinet J, Johnson S, and Pini G (2009). Applying generalizability theory using EduG: Quantitative methodology series. Routledge, New York, USA.

Clauser BE, Swanson DB, and Clyman SG (1999). A comparison of the generalizability of scores produced by expert raters and automated scoring systems. Applied Measurement in Education, 12(3): 281-299.

Dezhi C, Bi YH, Xitao F, and Kejian L (2014). Measurement quality of the Chinese early childhood program rating scale: An investigation using multivariate generalizability theory. Journal of Psychoeducational Assessment, 32(3): 236-248.

Fan CH and Hansmann PR (2015). Applying generalizability theory for making quantitative RTI progress-monitoring decisions. Assessment for Effective Intervention, 40(4): 205-215.

Gardner H (1993). Multiple intelligences: The theory in practice. Basic Book, New York, USA.

Harrington HL, Meisels SJ, McMahon P, Dichtelmiller ML, and Jablon JR (1997). Observing, documenting, and assessing learning: The work sampling system handbook for teacher educators. Rebus, Michigan, USA.

Hills TW (1993). Assessment in context: Teachers and children at work. Young Children, 48(5): 20-28.

Kleinert H, Greene P, and Harte M (2002). Creating and using meaningful alternative assessments. Teaching Exceptional Children, 34(4): 40-47.

Nancy JR (2001). Using authentic assessment to document the emerging literacy skills of young children. Childhood Education, 78(2): 66-69.

Ralph B and Geoffrey N (2012). Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. Medical Teacher, 34(11): 960-992.

Shavelson R and Webb N (1991). Generalizability theory: A primer. SAGE, California, USA.

Wehlage GG, Newmann FM and Secada WG (1996). Standards for authentic achievement and pedagogy. In: Fred MN (Ed.), Authentic Achievement: Restructuring Schools for Intellectual Quality: 21-48. Jossey-Bass, San Francisco, USA.