Studying Reliability of Open Ended Mathematics Items According to the Classical Test Theory and Generalizability Theory

Neşe GÜLER, Selahattin GELBAL***

Abstract

In this study, the Classical test theory and generalizability theory were used for determination to reliability of scores obtained from measurement tool of mathematics success. 24 open-ended mathematics question of the TIMSS-1999 was applied to 203 students in 2007-spring semester. Internal consistency of scores was found as 0.92. For determination of interrater consistency, Kendall's concordance coefficient was calculated as 0.52. Generalizability coefficient for mathematics scores was 0.92 and phi coefficient was 0.90. The variance component of raters accounted for 2.1% of the total variance. According to all results, it was seen that measurement tool of mathematics success was reliable for determination of students' mathematics success. Although there was a difference between means of four raters' scores, it was found that there was consistency of their scores.

Key Words

Reliability, the Classical Test Theory, Generalizability Theory, G-Study and D-Study.

* *Correspondence:* Assist Prof. Neşe GÜLER, Sakarya University, Faculty of Education, Department of Educational Sciences, 54300 Hendek, Sakarya/Turkey.
E-mail: gnguler@gmail.com
** Assoc. Prof. Selahattin GELBAL, Hacettepe University, Faculty of Education, Department of Educational Sciences, 06532 Beytepe, Ankara/Turkey.

Kuram ve Uygulamada Eğitim Bilimleri / Educational Sciences: Theory & Practice
10 (2) • Spring 2010 • 1011-1019

Being able to make important and correct decisions in the educational process, as in other branches of science, depends on reliable and valid measurement which is described as the results of any observation that are matched with numbers or other symbols (Baykul, 2000). Therefore, the reliability and validity study is one of the most emphasized issues by researchers. In such studies, which are performed in the fields of psychology and education, the Classical Test Theory based on the assumption that the observed score is a combination of the true score and error which is unique and cannot be separated usually forms the foundation (Brennan, 1992). As different from the term of error in this assumption, the Generalizability Theory- in which reliability study is possible by considering more than one source of error at the same time and which is actually the extended and flexible form of the Classical Test Theory- has been available recently for reliability and validity studies (Güler, 2009). Although the Generalizability Theory is originally based on the Classical Test Theory (CTT) and analysis of variance (ANOVA), it should not be considered as a simple combination of them (Brenman, 1992). Considering that the CTT and ANOVA are the parents of the Generalizability Theory, It has a similar and as well as a different structure rather than being just a simple combination of the two (Goodwin, 2001).

The Classical Test Theory

Reliability coefficient is defined as the proportion of true score variance to the observed score variance in the Classical Test Theory. The observed score variance is composed of two different sources: True score variance (which is thought to be systematic) and error variance (which is thought to be random). Reliability, which is usually described as the consistency of scores obtained through measurement may differ according to the source of error and thus is called accordingly. For instance, the correlation between scores which are obtained by two or more implementations of a test (test-retest) is called "stability," and the source of error here is thought to be the time between applications. Yet, in this sense of reliability, variance stemming from items sample is not taken into consideration (Güler, 2008).

The extent to which different raters give the same scores independent of each other to each student is called *interrater reliability*. Interrater reliability is commonly calculated with correlation coefficient, the degree of linear relations between the two raters. The correlation coefficient, which is called Pearson Product-Moment Correlation Coefficient (r), is found as the percentage of true variance in the total variance. Here, the remaining percentage means error representing the inconsistency between scores given by the raters (Anastasi & Urbina, 1997). A comparison of score averages along with correlation coefficients is also appropriate in calculating the interrater reliability. In case of only two raters it would be a better solution to interpret the score averages for matched groups through t test but in case of more than two raters a comparison with ANOVA for repetitive measurements and interpretation with correlation coefficients would be better (Goodwin, 2001).

When more than two raters rate a certain behavior, performance, question, etc; one of the methods employed in determining interrater reliability is Kendall's Concordance coefficient, a non-parametric statistical technique (Cooper, 1997; Howell, 2002).

The Generalizability Theory

The Generalizability Theory is an extension of the Classical Test Theory (Cronbach, Gleser, Nanda, & Rajaratman, 1972). The Classical Test Theory is based on the assumption that each observation or test score having one single true score generates one single reliability coefficient for a group of parallel observations. This assumption may be reasonable when the parallel forms are equalized carefully whereas it is not realistic or is restrictive when average scores or variances are different or when items on the form are heterogeneous.

Reliability in the sense of internal consistency tends to be low in a multi-dimensional measurement; but at the same time, test-retest and parallel forms reliability can be high. This case shows the restrictions and contradictions of the Classical reliability model. The Generalizability Theory has been developed as a more flexible method for removing the restrictions; and it enables to analyze errors likely to originate from potential sources of variance such as the rater, time, measurement form, tasks or items altogether and simultaneously and to calculate a comprehensive reliability coefficient. Also the Generalizability Theory eliminates the traditional difference between reliability and validity (Allal & Cardinet, 1997).

The achievement a student demonstrates in a task to be measured is considered as a sample drawn from a complex population in which all the potential sources of variance such as the rater and the task are available side by side. According to Shavelson and Webb (1991), the Generalizability Theory is believed to be the extended form of the Classical Test Theory in four respects: 1) The Generalizability Theory handles multiple sources of variance through one single analysis; 2) it assures determining the size of each source of variance; 3) it enables the calculation of two different reliability coefficients concerning making the relative decisions for individuals' achievements and the absolute decisions for individuals' achievements; 4) it enables to arrange measurements in which measurement errors are reduced to the minimum depending on the purpose (D-studies). In brief, the Generalizability Theory is an appropriate theory to predict reliability in measuring achievement.

Such factors as items or tasks, time and rater are called the *source of variance (facet)* in the Generalizability Theory (Brennan, 2001). In other words, facet is a term used for all sources of potential measurement errors. Thus, minimizing the rate of variance related with the source of variance as far as possible rather than maximizing it is preferable (Alharby, 2006). The levels of sources of variance are called the *conditions*. For instance, given that 'raters' are a source of variance; the first, second, third, etc rater each is a condition. The potential conditions of a source of variance which is usually available are considered to be infinite in size. All the potential observations to replace the current sample for which observations are made are called "*the universe of admissible observation*". The *"Universe of generalization"*, on the other hand is the set of conditions the researcher would like to generalize. In other words, individuals serve as the targets of the measurement for which decisions will be made. Therefore, individuals are not considered as a source of variance since variance depending on individuals is a desirable case. *Universe score* is defined as a measurement score which is found as the average of scores obtained from the universe of admissible observations for the sources of variance. Universe scores variance is similar to the true score variance available in the Classical Test Theory; yet as different from it, two different variances of error are available in the Generalizability Theory. The difference stems from the fact that it is possible to make decisions in two senses in the Generalizability Theory. Both the relative and the absolute variances of error are available in the Generalizability Theory. The relative error variance here may be thought like the error variance in the Classical Theory (Shavelson & Webb, 1991).

In addition to the basic concepts and terms mentioned above, different studies and designs are also available in the Generalizability Theory. Depending on the purpose of the research, they may be stated as G and D studies, source of variance to be random or fixed, design in obtaining the data to be cross or nested, interpretation of results to be relative or absolute, and as possibility of calculating the generalizability (G) coefficient or reliability coefficient (Phi).

It is believed in this current research that by testing the consistency, contributions may be made in predicting the reliability based on the Classical Test Theory and on Generalizability Theory- which have been suggested for similar measurement situations. When they employed subjective tools of measurement, mathematics educators wish to know how reliable the scores given in relation to mathematics knowledge are. Besides that, they also need to know what the most effective source of variance is in measuring mathematical achievement and how the measurement should be done to minimize measurement errors. Therefore, discussing those points and illuminating mathematics educators in those respects are the second purpose of this research.

Method

The study of the research is composed of 203 8th and 9th graders of different locations and school types in Ankara. Since the application concerning the measurement of mathematical achievement was conducted in class hours, the main element determining the student selection was the teacher's and the students' willingness. 24 open-ended items in the TIMSS were applied to the students. The items were assigned scores by using a holistic rubric in a manner so as to assign 0 to no answer, and 5 to the sample answer (Wiggins, 1998). Since 6 items were seen to have high load factor following the paraphrasing and confirming factor analyses, decision was made to remove them (Büyüköztürk, 2006).

All the analyses were conducted for the remaining 18 items. Four mathematics experts who are able to measure mathematical achievement took part in the research. Those raters performed their rating independently of each other.

The data analyses were conducted in two stages. In the first stage, factor structures which were determined through paraphrasing factor analysis were tested through confirmatory factor analysis. Paraphrasing factor analyses were done for the scores obtained from the four raters, and were seen to generate similar outcomes. However, since the descriptive statistics of the entire raters differed, paraphrasing and confirmatory factor analysis results for the average scores given by the raters to each item were included. In the final stage, the data were analyzed through designs a compatible with the Classical Test Theory and Generalizability Theory, respectively, according to the sub-problems of the research. In Cronbach Alpha, paraphrasing factor analysis and in calculating Kendall's Concordance Coefficient, SPSS (14.0) program was used. Lisrel 8.7 program was employed for confirmatory factor analysis. According to the Generalizability Theory, to predict the variance values for the main and shared effects and to calculate the G and Φ (Phi) coefficients for the reliability of scores, SPSS (14.0) program was used (Musquash & O'Connor, 2006).

Results

The reliability of the scores was estimated for each rater separately. For this purpose, Cronbach alpha coefficients were calculated. The reliability coefficients of the scores for first and second raters and for third and fourth raters were the same and found as 0.91 and 0.92, respectively. According to these results, it is predicted that the items measured mathematic success consistently. The consistency level of the four raters' scores was analyzed with Kendall's Concordance Coefficient and this coefficient was found to be 0.52 (χ^2 =315.16, sd=3, $p<.05$). Besides, the correlation coefficients for each pairs of rater scores were calculated and found the values which change between 0.90 and 0.97. These values support that there is a consistency between the raters' scores. The difference of the mean scores of four raters was analyzed with test of dependent samples analysis of variance and it was found that there is a statistical significant difference between mean scores (F=13.801, $p<.05$). After this result was found pairwise comparisons for means of raters' scores was examined by post-hoc studies. According to post-hoc studies it was found that there was not statistical significant difference between all paired raters' mean scores, except first rater.

For the purpose of estimating variances and percent of variances in G study, fully crossed (bxgxp) design, in which the responses of every student to all items are each scored by all raters, was applied. After examining the estimating variances and percentages of total variances it is seen that the variance component for students (b), which indicates the variance for a student mean score over raters and items, accounts for 32.6 % of the total variance in the scores. This can be interpreted that students systematically differed in their level of mathematic success. The variance component for students is the second large component in all. In generalizability studies, variance due to students is considered as a universe score and this variance shows that the difference between students in terms of characteristic which was measured (Brennan, 2001; Shavelson & Webb, 1991). The fact that the largest variance component belongs to students is wanted in the Generalizability Theory. Variance due to items (g) accounts for only 4.5% of the total variance which suggests that only less than 5% of the variability found in the model accounts for difference among items. The variance component for raters (p), which is the lowest variance components in all main variances, accounts for 2.1% of the total variance. This indicates that all raters are same level of lenient/severe across all students. Thus it can be said that there is a consistency between raters' scores.

The largest variance component is that associated with the interaction between students and items (bg) which accounts for 43.6 % of the total variance. This indicates that items differed in difficulty level; some items were easier than others. Variance due to interaction between students and raters (bp), which is the second lowest variance components across all variances, accounts for 1.3% of the total variance. This can be interpreted that all raters are almost same level of lenient/severe for all students. The smallest source of variation in this model is due to items by rater interaction which accounts for %0.7 only of the total variance. This suggests that all raters scored items almost same across the students. A third large component, residual error, indicates a large student by items by rater interaction, unmeasured sources of variation, or both. This variance component which represents the unwanted variance accounts for 15.3% of the total variance.

According to 18 items and 4 raters, G and Φ coefficients were estimated 0.917 and 0.897, respectively. Based on variance component estimates found in G study, different D-study scenarios were investigated. The first D-study conducted was one that assumes there were 18 items, which were same in the G-study, G and Φ coefficients were found as 0.912 and 0.888, respectively. In second D-study scenario when there were same number of items and there were 5 raters, G and Φ coefficients were found as 0.919 and 0.902, respectively. In other D-study scenarios, in this time, the number of raters is same and the number of items was changed. When it was assumed that there were 12 items G and *Φ* coefficients were estimated 0.884 and 0.863, respectively and when it was thought that there were 24 items then G and *Φ* coefficients were found as 0.934 and 0.915, respectively. Because the variance due to raters accounts for very small percentage of the total variance (0.075%), that the number of raters is increased or decreased does not much more affects the G and *Φ* coefficients. Thus it can be said that for both G and *Φ* coefficients, that the number of items is increased should be more effective than that the number of the raters is increased.

Because the Generalizability Theory eliminates the conventional difference between reliability and validity, Generalizability Theory by which reliability and validity can be determined with only one study should be suggested as an economic and practical theory. With different D studies reliability for different conditions which include different number of items, raters, etc. than original study can be estimated. It can be suggested that new studies which there are different designs (e.g. nested or mixed) or different facets are applied. When considered advantages of the theories, especially if the reliability is estimated when there are more than one rater, it can be said that besides the Classical Test Theory, using Generalizability Theory should be more explanatory and informative. In different measurement conditions, it can be thought that similar studies which both theories are examined together provide supplement to reliability studies.

References / Kaynakça

Alharby, E. R. (2006). *A comparison between two scoring methods, holistic vs. analytic using two measurement models, the generalizability theory and the many facet rasch measurement within the context of performance asssessment.* Unpublished doctoral dissertation, The Pennsylvenia State University, USA.

Allal, L., & Cardinet, J. (1997). Generalizability theory. In J. P. Keeves (Ed.), *Educational research methodology and measurement an ınternational handbook* (2nd ed., pp. 734-741). Cambridge University Press.

Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ.: Prentice Hall.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Ankara: ÖSYM.

Brennan, R. L. (2001). *Generalizability theory*. Iowa City, IA: ACT Publications.

Büyüköztürk, Ş. (2006). *Sosyal bilimler için veri analizi el kitabı, İstatistik, araştırma deseni, SPSS uygulamaları ve yorumu* (6. baskı). Ankara: PegemA Yayıncılık.

Cooper, M. (1997). Nonparametric and distribution-free statistics. In J. P. Keeves (Ed.), *Educational research methodology and measurement an ınternational handbook* (2nd ed., pp. 607-612). Cambridge University Press.

Cronbach, J. L., Gleser, G. C., Nanda, H., & Rajaratman, N. (1972*). The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley and Sons.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Psychical Education and Exercises Science*, *5*(1), 13-14.

Güler, N. (2008). *Klasik test kuramı genellenebilirlik kuramı ve Rasch modeli üzerine bir araştırma.* Yayımlanmamış doktora tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Güler, N. (2009). Genellenebilirlik kuramı ve SPSS ile GENOVA programlarıyla hesaplanan G ve K çalışmalarına ilişkin sonuçların karşılaştırılması. *Eğitim ve Bilim*, *34*, 154.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Thomson Learning Academic Research Center.

Mushquash, C. & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analysis. *Behavior Research Methods*, *38*, 542-547.

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

Sudweeks, R. R., Reeve, S. & Bradshaw, W. S. (2005). A comparison of generalizability theory and many facet measurement in an anlaysis of college sophomore writing. *Assessing Writing*. 9, 239-261.

Wiggins, G. (1998). *Educative assessment: Designing assessment to inform and improve student performance*. San Francisco: Jassey-Bass Publishers.